

After the gold-rush

Wall Street's technology-led boom distorted the priorities of research agencies. Its bust should remind scientists that public trust requires careful attention to ethics.

The investment bank Morgan Stanley is closing down its Competitive Edge Best Ideas mutual fund, reports *The New York Times* in one of those stolid and yet quietly humorous stories that have recently characterized the financial pages. Apparently the 'best ideas' of Morgan Stanley's analysts weren't coming to much, and the fund had lost half of its value in two years. That's enough bright ideas for the moment, investors seem to think.

Not all of the bright ideas touted on the world's stock markets towards the end of the 1990s were grounded in science or technology. Some, such as persuading people to buy dog food over the Internet instead of at a supermarket, were more like the product of a students' drinking party.

However, the boom was led by so-called technology stocks — and was fuelled by sometimes unrealistic assumptions about the effect of science and technology on long-term economic growth. At the height of the boom, then US President Bill Clinton and other world leaders accredited science and technology with creating a 'new economy' that somehow justified the wanton overvaluation of stocks, especially technology stocks. Although most of these stocks were associated with information technology and telecommunications, the publication of the human genome and even the much-vaunted promise of nanotechnology were associated with economic well-being in general and some high-flying bubble stocks in particular.

That was then. Today, a more realistic set of investor expectations about the commercial potential of innovation are in the ascendant. This in turn is already reducing the number of commercial opportunities that are available to academic researchers, even in such hotbeds of technology transfer as San Francisco and Boston. For science as a whole, such realism is no bad thing, especially if it also leads to a correction of a short-sighted historical trend.

Commercial interest

During the economic boom of the 1990s, the connections between science and business deepened and widened. Despite a drumbeat of protest from 'traditional' academics — the very epithet reflects their marginalization — this trend was getting out of hand. Commercial interests were not just present on many university campuses, but paramount. The problem was less pronounced at such places as Harvard and Berkeley, which could afford to name their terms and their price, than at some of the middle-ranking universities that aspired to compete with them. This happened not only across the United States but around the world.

The price that the research universities paid was one of balance, as resources flooded into disciplines of commercial interest. Researchers in cell biology and information technology built palaces of gold, while purveyors of other branches of knowledge — whole-organism biology and archaeology, for example, and even mathematics — have been shunted to the periphery, at least in the eyes of university administrators.

The same pattern has manifested itself in the funding priorities of research agencies, in the United States and elsewhere. Although it is in the long-term interest of any nation to develop a well-balanced

portfolio of investment in curiosity-driven science — of the sort articulated most famously by Vannevar Bush for the US National Science Foundation at the end of the Second World War — few have done so, especially of late. Put simply, no one has been at the table to stand up for anthropology or geology, say, as politicians have looked to science primarily as an instrument to fuel growth in a few fashionable industries. During the past decade, in nation after nation, large grants and new projects have been focused on biotechnology, information technology and, more recently and to a lesser extent, nanotechnology.

Like Wall Street, the political leadership of research agencies has been overly fixated by these mantras. And like Wall Street, it should now rediscover the importance of what the money men call 'fundamentals'. The fundamental goal of science policy should be a well-balanced system of universities and government laboratories that explore the frontiers of knowledge. Surveys of public opinion in the United States show solid public support for investment in basic research that does not promise short-term economic benefits.

A matter of trust

The technology boom and bust may have distorted scientific priorities but it has not done much damage to the integrity of universities or to scientific institutions. At a time when scandals are shredding so many reputations — with businesspeople and accountants joining politicians and journalists as targets of public opprobrium in the United States — trust in science remains deservedly high.

The scandals that have shaken financial markets have touched on science, however. For instance, ImClone Systems, one of the companies where allegations of insider dealing have shaken investor confidence, is a biotechnology company with friends in high places in US biomedical research. But so far science as a whole has escaped damage. Nevertheless, dangers lurk as remaining scandals unfold: the situation at Lucent Technologies, for example, continues to warrant careful attention (see *Nature* **418**, 5; 2002). And complacency about scientific misconduct still pervades the community to a surprising degree (see pages 113 and 120).

The possible cost of such complacency is, perhaps, the most important message that scientists can draw from this year's financial scandals. It is tempting but misguided for scientific leaders to believe that, because a false result will be corrected in time by its failure to be replicated, research misconduct doesn't really matter. Business leaders had a similar credo: they said that markets would correct for dishonest business practices, in the fullness of time.

This isn't good enough. Perhaps a better working assumption should be that people will get away with what they can get away with. Since the obdurate Congressman John Dingell (Democrat, Michigan) was deposed as chair of the Commerce Committee of the US House of Representatives in 1994, political scrutiny of the scientific community over misconduct has relaxed. Most researchers are honest and not wealthy, and are not surprised to learn that corporate America is full of rogues. Before drawing too much comfort from their come-uppance, however, it would make sense to ensure that research has its own house in order. ■





Identity parade
Europe backs
extensive labelling
of transgenic foods
p114



Animal harm
The push is on for
animal-free toxicity
tests in Europe
p116



Production line
Physicists draw up
plans for neutrino
factory
p117



Tribal counsel
Anthropology
society releases
Yanomami report
p118

Dubious data remain in print two years after misconduct inquiry

Alison Abbott and Johanna Schwarz, Munich

Many of the 94 scientific papers listed two years ago by a German inquiry as likely to contain manipulated data have yet to be retracted from the literature, *Nature* has discovered.

The suspect papers were identified during a high-profile investigation into misconduct in German cancer research, the results of which were published in June 2000.

Nature surveyed 29 journals asking about their response to the inquiry's findings. Twenty replied, which between them published 60 of the papers. Five said that they had retracted all of the papers (22 in total), one had retracted one of three it had published, and 14 had not retracted any. Among the 20 journals, seven said they were unaware that the investigation had taken place, and four more did not know it had been completed.

Scientific leaders and misconduct investigators around the world have long complained that, even when scientific misconduct is proven, no reliable mechanisms exist to remove bad information from the literature.

"Somebody must be responsible for informing the journals if papers have been found to contain wrong data," says Gerhard Neuweiler, a zoologist at the University of Munich and former president of Germany's science council. "And the journals must find ways to alert the community. Something must happen, because we can expect misconduct to happen more often in the future

— in particular in biomedicine, where the pressure to publish is very high."

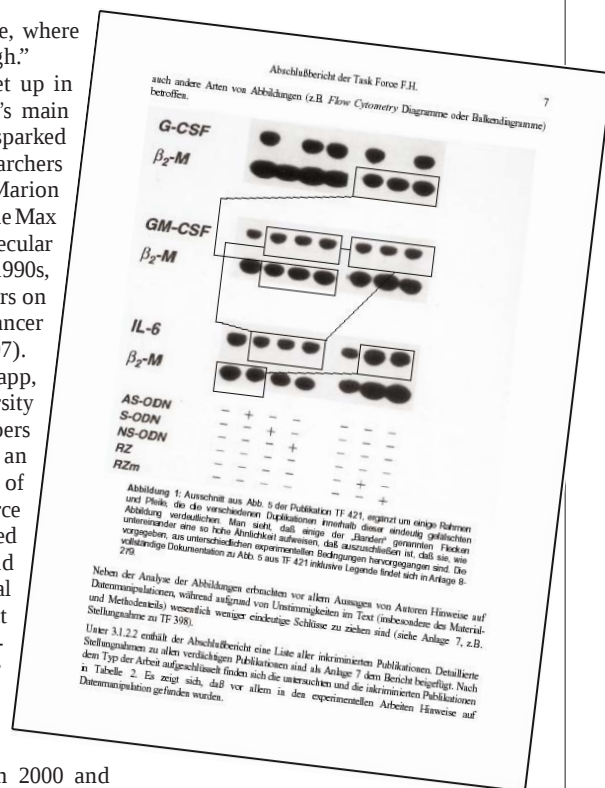
The German inquiry was set up in 1998 by the DFG, the country's main research funding agency. It was sparked by allegations that cancer researchers Friedhelm Herrmann and Marion Brach, who worked together at the Max Delbrück Centre for Molecular Medicine in Berlin in the early 1990s, had fabricated data in four papers on drug-resistance genes and cancer therapy (see *Nature* **387**, 442; 1997).

A task force headed by Ulf Rapp, a cell biologist from the University of Würzburg, examined 347 papers co-authored by the pair in an attempt to determine the extent of the fabrication. The task force wrote to all of the journals involved explaining the situation, and requesting records of the original submissions. It concluded that data in 85 papers and 9 book articles had either definitely or "highly probably" been manipulated. A German-language report, complete with a list of suspect papers, was released in 2000 and published on the DFG's website.

But the journals involved were not officially notified of the findings. The main aim of the investigation was not to clean up the literature, DFG officials explain, but to make sure that junior scientists whose names were on the papers could be identified and exonerated.

Rapp, however, says he understood that the task force's work would also help to set the published record straight. He now accuses the DFG of failing to do this by not informing the journals of the inquiry's conclusions. "I reminded the DFG on multiple occasions," he says.

Some editors, too, were expecting to hear about the outcome of the investigation. Barbara Cohen, executive editor of the *Journal of Clinical Investigation*, says that the original task-force letter sent in May 1998 promised this. "We are still waiting for the follow-up," she says. A spokesman for *Acta Haematologica* says that, having heard nothing further from



Data duplication reported by the DFG — this paper has been retracted but others identified as suspicious are still in print.

the task force, "we assumed that the papers were freed from suspicion".

Other editors do not remember the original letter. Mike Kemeny of King's College London, for example, took over as editor-in-chief of *Immunology* two years ago. Alerted by *Nature*'s survey, Kemeny says that he will now "take the issue forward".

Nature's survey also indicates that journals lack a common policy for retracting papers. Although most require authors to request retraction, a few editors were prepared to act independently in the Herrmann and Brach case — partly because Herrmann has been advised by his lawyer not to retract any of his papers.

► www.dfg.de/aktuell/download/abschlussbericht_fh.pdf





Shelf help: European supermarkets may soon label all foods containing transgenic material.

Europe gets tough on labelling genetically modified foodstuffs

Declan Butler, Paris

Clear labelling of all foods containing material from transgenic organisms may be coming to Europe's supermarkets soon, after the European Parliament backed a proposal to implement it.

The plan, which could be in place within two years, has angered the food industry, which claims that stricter labels would infer a health risk that has not been proved to exist. It could also provoke more trade friction between the European Union and the United States, which strongly opposes such labelling.

On 3 July, the parliament supported a European Commission proposal to increase labelling requirements for foods. It also backed tougher rules than the Commission suggested to test and identify foods and animal feed that have been contaminated with genetically modified material during production or transportation.

The parliament also supported rules that will require manufacturers to trace foods to their place of origin. It narrowly rejected the compulsory labelling of meat, milk and eggs obtained from animals raised on genetically modified feed.

But the parliament voted to cut the threshold above which foods contaminated with transgenic material are classified as genetically modified from 1% of content to 0.5%. Representatives of the biotechnology industry branded this plan 'unrealistic'.

EuropaBio, an industry group, said in a statement that the parliament's actions would "discriminate against the new technology, reduce consumer choice, disrupt trade with Third World countries, while adding nothing to safety".

But Rachel Sutton of the Consumers' Association in London welcomes the decision. She says that the threshold level for labelling should be continually reduced as testing methods improve. ■

Botswana's AIDS laboratory squares up to HIV pandemic

Michael Cherry, Gaborone

With one in three of its adult population infected, Botswana has the highest prevalence of HIV in the world. But in the dusty streets of the capital, Gaborone, there is still hope that the epidemic can be contained.

At the city's Princess Marina Hospital, for example, a dedicated team of 90 scientists and technicians are establishing Africa's largest single laboratory dedicated to AIDS research.

The lab opened its doors at the end of last year, the culmination of a long collaboration between Botswana's government and the Harvard AIDS Institute in the United States.

Most global AIDS research has focused on the HIV-1B virus found in the United States and Europe. So one question being explored by the Botswana lab is whether the HIV-1C virus encountered in Africa has traits that have contributed to its rapid spread through the continent. Researchers in Gaborone now say that more is known about HIV-1C subtypes in Botswana than anywhere else.

The laboratory's research has two main objectives: to determine the long-term effects of drug treatment, and to contribute towards the development of an effective vaccine for the virus. The latter aim, in particular, requires the molecular characterization of the strain and a better understanding of the evolutionary development of its 50-plus variants that have been identified in Botswana alone.

To aid the design of a subtype-specific HIV vaccine, cytolytic T-lymphocyte responses and human leukocyte antigen class specificities have been obtained from patients all over the country. Mapping these should provide an indication of the role of genetics in patients' responses to the virus and contribute to vaccine development.

It could also lead to an understanding of whether the pandemic's rapid spread in southern Africa is related to any property of the subtype itself. "It's difficult to tease out the evidence," says Trevor Peter, the lab's manager, "but it appears both to have a higher replicative rate and to be more transmissible,



Trevor Peter: on the lookout for evasive evidence.

possibly on account of having higher concentrations in body fluids."

Botswana, with a population of only 1.7 million, has no medical school, and most of the lab's senior researchers—including 10 of its 12 clinicians—are foreign. Half of the staff, however, are Botswanan.

Most of the staff are paid on grant money awarded through Harvard,



Full house: Gaborone's Princess Marina Hospital is home to a 90-strong team researching AIDS.

and salaries have not hindered recruitment, according to Max Essex, an AIDS researcher at Harvard who was heavily involved in its establishment. The Botswanan government paid about \$3.5 million to build the laboratory, and charities and drug companies contributed a similar amount to equip it.

Essex says that processing samples locally gives the lab a time advantage over researchers outside Africa, although materials needed to do the work can be slow to arrive.

Hermann Bussmann, who coordinates the lab's research on the impact of anti-retroviral drug treatment of HIV in Botswana, wants to know the extent to which poor adherence to complex anti-retroviral drug regimens will foster resistant strains of the virus. "Resistance will develop," he says, "but hopefully anti-retrovirals will buy enough time for a vaccine."

"The partnership between Harvard and the Botswana government is a great model," says Bussmann, who moved to the lab from Germany and is pleased to be "fighting the pandemic in the region where it is most needed". He firmly believes that the fight can be won. "Things may go slowly in Botswana, but at least they go forwards," he says. ■



On the front line: Hermann Bussmann.

AIDS meeting demands more than lip service

Declan Butler, Barcelona

The top priority for AIDS researchers and activists is to press governments to translate words into actions, the 14th International AIDS Conference in Barcelona has been told.

The last such conference — in Durban, South Africa, two years ago — generated many promises of action, but most are so far unkept. Opening the Barcelona meeting on 7 July, Peter Piot, executive director of UNAIDS, nonetheless said that it heralded “a new era — the era of AIDS as a global political issue”.

Some progress has been made since Durban. Spending on AIDS in poor countries has gone up, and many of these nations have established national strategies to fight the disease. Some countries that have implemented prevention and care programmes — notably Brazil and South Africa — have reversed the growth in HIV infection rates in their populations.

Drug companies have started making anti-retroviral treatments for AIDS more affordable — although these are still only available to a small number of patients in poor countries. And at the UN General Assembly Special Session on HIV/AIDS in June last year, governments set targets to reduce the spread of the disease and pledged to pay the US\$10 billion needed each year to fight the pandemic. Thus far, however, only \$2 billion of that has actually materialized.

Among positive developments so far this year, the release of the first grants from the Global Fund to Fight AIDS, Tuberculosis and Malaria (see *Nature* **416**, 773–774; 2002) will support work on prevention and drug treatment. The grants aim to increase the availability of anti-retroviral drugs in developing countries. UNAIDS estimates that only 30,000 people in Africa currently have access to the drugs. The price of annual anti-retroviral drug treatments has come down to US\$300 in some cases, but Médecins sans Frontières, the Paris-based medical aid charity, said that this must fall to \$50.

The global fund's work is expected to make the drugs available to 550,000 people in poor countries by 2007 — still a small proportion of the 40 million people, 94% of them in developing nations, who carry the disease today. Three million babies are being born each year with HIV, Piot said.

Bernhard Schwartlander, director of the Department of HIV/AIDS at the World Health Organization, told the conference that by current projections there could be 45 million new cases of HIV/AIDS by 2010. If governments comply with the promises made to the United Nations, 29 million of these could be avoided, he said, adding that each year's delay in implementing the promises will cost 5 million lives.

In some African countries, prevalence of

HIV runs at over 30%, and is as much as 50% in some cities. Without urgent action, this pattern could be reproduced in parts of Asia, the Middle East and Eastern Europe, according to UNAIDS projections.

Prevention is becoming even more central to containing the pandemic, as most scientists now agree that anti-retroviral therapy will not eradicate the virus from the body.

“The fact that the infection is intrinsically incurable with retroviral therapy is as power-

ful an argument for prevention methods as we will ever find,” Robert Siliciano, an AIDS researcher at Johns Hopkins University in Baltimore, Maryland, told the conference.

Piot said that by the time the next conference takes place in Bangkok in 2004, it will be easy to tell if governments have met their promises or not. “We will know who has delivered,” he said. “Bangkok will be a time of accountability.”

www.aids2002.com



Promising start: the AIDS conference in Barcelona opened this week amid calls for government action.

Biology project left short of cash

David Cyranoski, Tokyo

Japan's partners in the Human Frontier Science Program (HFSP) — which supports international projects that take a multidisciplinary approach to problems in biology — have rebutted its plea for a firm commitment to more cash for the scheme.

At the programme's annual conference last month in Berlin, the partners, which include the United States and several European nations, declined to translate their support for HFSP projects into a formal agreement.

The HFSP was initiated by the Japanese government in 1989 to support biological research involving international teams, most of whose members work abroad on fellowships. Although it targets biological problems, recent emphasis has been placed on projects involving chemists, physicists, mathematicians and engineers.

But the HFSP has suffered as the dollar value of Japan's contribution, which accounts for more than half of its total funding, has declined with the yen. The United States and Italy, in particular, also fell far short of their

targets. As a result, the HFSP has only US\$50 million in funding this year against a \$60-million target agreed by the partners in 1997.

According to one representative at the meeting, US officials angered other members at the conference by insisting that the 1997 target was non-binding. Research-agency representatives at the meeting agreed that the programme needs \$30 million annually from outside Japan by 2004 — but failed to agree on how to divide this up.

“We could not reach an agreement on a framework that would make this compulsory,” says Takayuki Shirao, the HFSP's deputy secretary. The meeting did agree to expand support for young scientists and to try to ensure that jobs are open for HFSP participants when they return home.

Japanese officials expect the programme to continue, but are disappointed that the partners are unwilling to increase their stake in it. “Everyone praises the HFSP's significance,” says Shirao. “But even after 10 years, the HFSP's tragedy is weak financial support for such international projects.”

Race is on to find alternative to animal tests

Sally Goodman, Paris

Plans to test systematically the toxicity of thousands of chemicals that are widely used in Europe have sparked a race to develop *in vitro* alternatives to tests involving animals.

Plans by the European Commission (EC) to test up to 30,000 substances by 2012 would stretch available animal-testing resources to the limit — and draw the ire of animal-rights protesters, observers say.

So researchers are calling on governments to use the screening plan as an incentive to ramp up the development of other tests that would use, for example, human skin cells to screen the substances in a test tube.

Currently, most toxicity testing is done *in vivo* using fish, mice or rats in contract research laboratories. By some estimates, 12 million extra animals would be needed for the testing proposed in an EC white paper, *Strategy for a Future Chemicals Policy*, that was published in February 2001.

But a report prepared by the European Centre for the Validation of Alternative Methods (ECVAM) in Ispra, Italy, says that with more funding for validation studies and infrastructure, alternative methods now being developed could replace animal tests, or at least greatly reduce the number of animals used in initial pre-screening of chemicals, by 2012. The report by ECVAM, which coordinates the validation of alternative testing methods in Europe, says that more research into areas such as the effects of toxicity on reproduction and of 'endocrine disruptors' could help to develop non-animal tests in areas that currently lack alternatives.

But some industrial researchers — many of whose employers oppose the EC's screening plan — are sceptical. "It is highly unlikely that we will be able to replace *in vivo* tests for the most complex effects, such as carcinogenicity or birth defects, within ten years," says Phil Botham, head of human safety at Syngenta's toxicology laboratory in Macclesfield, UK. A timescale of up to 30 years would be more realistic, he says.

"Pouring new money into *in vitro* research is not the solution," Botham says. "Much of the current *in vitro* research is done by researchers who don't understand toxicology," he adds, calling for the European Union (EU) to accept more guidance from industry on the sort of toxicological research that it ought to support.

Michael Balls, the retired head of ECVAM who edited the report, says that efforts to get alternative methods approved — which typically, he says, takes six years — are being hampered by a lack of EU funds for validating the methods.

Hella Lichtenberg, a biologist at the University of Bonn in Germany and coordinator of a consortium that is developing yeast



Testing time: plans to screen 30,000 chemicals for toxicity may require millions more animals.

expression technology as a non-animal method for pharmaceutical and toxicological screening, agrees that the EU should be funding projects for longer, so that they can be validated. "The time-consuming validation process requires additional funding to ensure that promising projects are kept running," she says.

The EC's proposal, which has yet to be accepted by EU governments, would aim to test the toxicity of industrial chemicals that were sold in bulk before 1981, when testing requirements for new chemicals came into force.

♦ <http://ihcp.jrc.cec.eu.int/Activities/ACTVali/ACTVali.html>

Japan lays out 55-point patent plan

David Cyranoski, Tokyo

The Japanese government has unveiled a plan to tighten controls on intellectual property — and encourage reluctant researchers to apply for patents.

The 55-point plan, which was published on 3 July by Prime Minister Junichiro Koizumi's Strategic Council on Intellectual Property, aims to establish graduate programmes to teach intellectual-property rights. It also offers the courts guidelines on the appropriate punishment for patent infringements, and a framework for implementing an agreement reached last month with the United States on mutual recognition of materials used in patent examinations. It also sets out rules for the transfer of tools and materials between researchers and research institutions.

"There are still many things to be done, but this is a milestone," says Takafumi Yamamoto, chief executive of the Center for Advanced Science and Technology Incubation, a company that patents inventions at the University of Tokyo. "This is the first sign of a clear direction for the use of intellectual property in Japan," he says.

Some say that the package will make it easier for researchers to patent their innovations and profit from them. Since 1999 Japanese law has allowed, but not required, funding agencies to yield patents to researchers or their institutions. Critics

say that many funding agencies have retained control of patents, but the new measures strongly encourage them to hand over control to research institutions.

The package broadly reflects proposals put forward by the National Forum for Intellectual Property Strategy earlier this year (see *Nature* 415, 354; 2002), although some elements are missing. "The proposals include provisions to strengthen education in patent law, but they don't include measures for bringing scientists into law classes," complains Koichi Sumikura, a specialist in intellectual property at the National Graduate Institute for Policy Studies in Tokyo and a member of the forum.

Some researchers fear that the government's obsession with intellectual property may lead it to fund science on the basis of patents, instead of quality. The reforms are scheduled to come into effect in 2004 as part of a shake-up of Japan's universities. Subsequently, the proposal document calls for a "distribution of resources to universities and public research institutions that appropriately reflects an evaluation of the institutions' progress in attaining and making use of patents".

"Rewarding patents is good, but not if it means a decrease in funding for basic science," says Keiichi Kodaira, president of the Graduate University for Advanced Studies in Kanagawa.

Physicists plot mass production of neutrinos

David Adam, London

The Large Hadron Collider may be years from operation, and the Next Linear Collider barely off the drawing-board, but ambitious high-energy physicists already have another project in mind.

Dozens of them met in London last week to lay plans for a neutrino factory that, they say, could shed much-needed light on the burgeoning field of neutrino physics.

Recent experiments have shown that neutrinos, which carry no charge, can oscillate between their different subtypes (see *Nature* **411**, 877; 2001), indicating that they have some mass. Such results have important consequences for the standard model, the theory that describes fundamental particles and their interactions, in which neutrinos had been assumed to have zero mass.

Physicists say that a neutrino factory would enable them to study neutrinos and their oscillations in far greater detail. Most observations so far have been of naturally occurring neutrinos striking the Earth from the Sun and from supernovae, and efforts are under way to study neutrinos produced in particle accelerators and nuclear reactors. But the proposed neutrino factory would produce a much more intense beam for researchers.

Funds are scarce, but Ken Long, a high-energy physicist at Imperial College, London, who organized the recent meeting, says that the project is making steady progress. It will take at least another five years to finalize the design, he says, but initial studies indicate that



Funds permitting, the Rutherford Appleton Laboratory will test elements needed for a neutrino factory.

the facility could be put into operation by 2015 at a cost of around £2 billion (US\$3 billion).

At the heart of any neutrino factory would be a high-power proton beam, which would crash into a solid target to yield an exotic concoction of particles. The factory would purify this mix and use particle accelerators to produce an intense neutrino beam.

To do this, researchers must isolate muons, unstable particles with a mass some 200 times that of an electron, which decay to form neutrinos. "The muons will have a lifespan of only about 2 microseconds," says neutrino researcher Michael Zisman of the Lawrence Berkeley National Laboratory in California. "But that's enough time for us to grab them, shape them, accelerate them and put them into a storage ring—or at least we think it is."

A £20-million project to test the 'muon

ionization cooling' concept needed to steer the muons is planned for the Rutherford Appleton Laboratory in Oxfordshire — if funding can be obtained.

Zisman says that a major technical challenge for a neutrino factory is designing a target to withstand the destructive proton beam. One idea is a carbon rod immersed in helium, to slow its physical deterioration. Another would use a jet of liquid mercury.

And some physicists view the neutrino factory as a stepping stone to yet another future project. They think that learning to control and store muons in a neutrino factory would open the way towards building a muon collider (see *Nature* **394**, 611; 1998), which could use the relatively massive particles to yield very high-energy collisions in a machine of a manageable size. ■

Public body appointed to clean up UK's nuclear legacy

David Adam, London

Britain took its first steps towards dealing with its stockpile of nuclear waste last week, when the government announced a new public body to clean up what it terms the country's "nuclear legacy".

The Liabilities Management Authority (LMA) will be responsible for treating and packaging waste from nuclear-fuel reprocessing and contaminated material from nuclear power stations, and for cleaning up their sites and the surrounding land.

But it won't tackle materials from the nuclear-weapons programme or consider the need for long-term waste disposal. Even so, the clean-up is expected to cost some £48 billion (US\$73 billion) over several decades.

Announcing the plan in a white paper on 4 July, energy minister Brian Wilson pledged that the new authority would implement the clean-up safely, securely and cost-effectively.

The LMA will take over responsibility for

the civil nuclear plants operated by the government-owned British Nuclear Fuels Limited (BNFL). But BNFL will continue to operate the sites, including the reprocessing plant at Sellafield in northwest England, under licence. By effectively taking the cost of waste treatment off BNFL's balance sheet, the move opens the door to the privatization of the company.

Geoffrey Boulton, vice-principal of the University of Edinburgh and chair of a Royal Society working party on the management of radioactive waste, cautiously welcomed the move. "In broad terms it's sensible because something has to be done, but it still needs to be meshed into a more complete solution," Boulton says.

"This is about how to get the waste into the tins, not what to do with the tins," says one official. A separate government consultative process addressing wider issues, including the long-term storage of waste, is expected to report this year.



Clean sweep: under the UK plans, a dedicated agency will be responsible for nuclear waste.

Britain's plans to store nuclear waste permanently were thrown into disarray in 1997, when research into a deep-storage facility at Sellafield collapsed after planning permission for an underground laboratory had been refused (see *Nature* **386**, 423; 1997). ■

Report pronounces on anthropology misconduct

San Diego The American Anthropological Association has issued its final report on controversial studies of the Yanomami people in Venezuela.

Two years ago, journalist Patrick Tierney published the book *Darkness in El Dorado: How Scientists and Journalists Devastated the*

Amazon, in which he alleged that researchers conducted clinical studies without informed consent and may have exposed the Yanomami to infectious diseases (see *Nature* **408**, 391; 2000). One of the accused, James Neel of the University of Michigan, died before the book was published. The other, Napoleon Chagnon, an emeritus professor of the University of California, Santa Barbara, has protested his innocence.

The association's report, issued on 2 July and made available online, confirms some of the charges. It says that Chagnon violated the association's ethical code in 1990 when he went into Yanomami territory after being denied a research permit, thereby endangering tribes because of inadequate disease-control measures. But the report stops short of backing Tierney's allegation that Neel's use of a potentially dangerous measles vaccine may have spread the disease among the Yanomami.

"The most critical thing we learned is that these people really are in terrible danger," says Jane Hill, a University of Arizona anthropologist who chaired the team that wrote the report.

♦ www.aaanet.org

German law accused of stifling foreign collaboration

Munich The head of the DFG, Germany's main research funding agency, has slammed the German government's recent ruling on the use of human embryonic stem cells, claiming that it is damaging collaborations with foreign scientists.

Ernst-Ludwig Winnacker, the DFG's president, said that the new legislation, which forbids German scientists from using stem cells while abroad (see *Nature*

417, 8; 2002), has become an obstacle to international cooperation. "German stem-cell researchers have for the time being put their planned collaborations with Israeli and British scientists on ice," Winnacker told the DFG's annual assembly, which was held last week in Berlin.

Together with a new ruling that gives animals a constitutional right to be protected by the state (see *Nature* 417, 372; 2002), the new regulations leave "a stale taste of hostility to research", says Winnacker. The DFG has commissioned a report to clarify the legal position of German scientists wanting to join international stem-cell research projects.

Safety risks of sewage dumping still murky

Washington The scientific case for the safe dumping of biosolids — sludge-like substances less euphemistically known as treated sewage — remains unclear, according to a new study from the US National Academy of Sciences.

Dumping of biosolids on land has increased since ocean dumping was banned in 1992. Some 3.4 million tonnes are now dumped every year in the United States, either as landfill or for use as fertilizer. Citizens' groups and environmental organizations have raised a stink over the



Yanomami: a people "in terrible danger".

practice, claiming that it spreads disease and poses other health risks.

The academy's panel, which released its report on 2 July, found no evidence that public health has been compromised by the land application of sludge and gave no opinion about whether the Environmental Protection Agency regulations covering the practice were adequate. Instead, it said the agency should assay chemicals and biological pathogens in the sludge and carry out epidemiological studies and risk assessments.

♦ www.nas.edu

Disease specialist will head top public-health centre

Washington After three months without a leader, one of the United States' main public-health agencies has a new hand at its helm: infectious-disease specialist Julie Gerberding.

Health and human services secretary Tommy Thompson named Gerberding as director of the Atlanta-based Centers for Disease Control and Prevention (CDC) on 3 July, in an announcement that was widely welcomed by public-health experts.

Gerberding, who has held various posts in the CDC since joining in 1998, takes over an agency that must define its role in combating bioterrorism, together with the FBI, the proposed Department of Homeland Security



Gerberding: willing team member.

"She is very competent and committed."

♦ www.cdc.gov

and the National Institutes of Health, while improving public-health infrastructure. "She is willing to pitch in as part of a team," says Tony Fauci, director of the National Institute of Allergy and Infectious Diseases in Bethesda, Maryland.

Supercomputer to aid African HIV research

Cape Town Africa's first research supercomputer will be switched on this week.

The computer is based at the University of the Western Cape's South African National Bioinformatics Institute near Cape Town. Researchers will use the machine — a Cray SV1 — to develop algorithms for analysing genetic information, particularly on HIV.

Winston Hide, the institute's director, says the supercomputer will be a major boost to South Africa's biotechnology infrastructure. It will also give researchers and students access to supercomputing time without having to book time on foreign machines.

♦ www.sanbi.ac.za

Prospectors invoke mystery particles to divine for oil

London An oil company working in Britain is planning to track emissions of subatomic particles in its search for black gold. But there appears to be a flaw in the plan: physicists say that the particles don't exist.

Technology Investment and Exploration, a company based on the island of Guernsey, has been given the go-ahead by the Department of Trade and Industry to begin tests for oil at three locations in England and Wales. The company plans to monitor the sites for emissions of 'microleptons' — tiny members of the lepton family of particles that the company claims are emitted by underground oil deposits.

But microleptons are missing from the physics textbooks. "There is no evidence for their existence," says Martin Perl, the Stanford University physicist that the company says developed the theory behind their technique, although Perl dissociates himself from their work. The particles were, however, the subject of Russian research in the 1980s, and the prospectors are not the first to believe in them. "If you search the web, you will find that some New Age philosophy or health sites have taken up the microlepton idea," says Perl. "There seems to be no end to human folly."

Time to wise up?

With one of their brightest young stars facing allegations of data falsification, Geoff Brumfiel asks whether researchers in the physical sciences are ready to tackle the issue of misconduct.

The corridors were buzzing with talk about Jan Hendrik Schön's work on nanoelectronics in March at the American Physical Society's meeting in Indianapolis, Indiana. The young researcher was a wizard in the field: he had produced some 80 papers in the previous two years. But other researchers had little luck reproducing his results, and were growing suspicious. Some suspected that Schön was withholding information to hinder competitors. Others were murmuring the unthinkable: could this be scientific fraud?

Then in May, sources inside Schön's workplace — Bell Laboratories in Murray Hill, New Jersey — revealed to an academic colleague that two graphs in separate papers seemed identical down to what should have been random noise (see *Nature* 417, 367–368; 2002). Since then, the list of suspect papers has steadily grown. Bell Labs convened an external panel to investigate; its ruling, on what is already the biggest case of alleged misconduct in the physical sciences, is expected by the end of the summer.

Regardless of the outcome, experts say the case has exposed a community poorly equipped to deal with misconduct. While biologists have been forced to address the issue following a series of high-profile investigations, misconduct seems to fall below the radar for researchers in the physical sciences. *Nature's* inquiries suggest that most are unaware of how to detect suspicious activity in their labs — or how to respond if allegations arise. If this deficit isn't addressed, say experts, the reputations of individuals, labs and even entire fields will remain at risk.

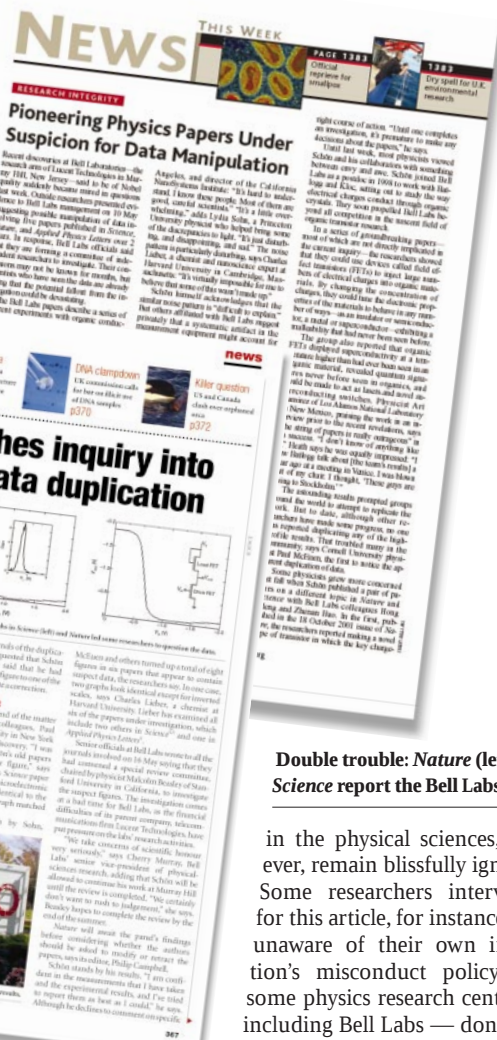
Scroll back two decades, and the same situation prevailed in US biomedicine. But that changed in the 1980s, when a Congressional committee headed by Representative John Dingell, a Michigan Democrat, started investigating allegations of misconduct relating to projects funded by the National Institutes of Health (NIH). Although many biologists regarded the exercise as a witch-hunt, it did eventually result in a system for dealing with alleged misconduct.

Today, all US universities receiving federal funding are supposed to have policies detailing how misconduct investigations should be conducted. At the federal level, the Office of Research Integrity (ORI), which operates from the NIH's parent body, the Department of Health and Human Services, reviews cases involving NIH grants after the researcher's institution has completed its investigation — and can debar miscreants from federal funding for a defined period.

The National Science Foundation (NSF) has the Office of the Inspector General, the agency's independent watchdog, performing a similar function. In December 2000, after much debate over the definition of misconduct — eventually narrowed to falsification, fabrication and plagiarism — the White House Office of Science and Technology Policy issued guidelines that formalized procedures for the whole of the US government.

In the dark

Few would claim that the system is perfect. Investigators often struggle to pursue important cases to fruition, and are sometimes hampered by legal action from the accused. But at least biologists know that the system is there. Many of their colleagues



Double trouble: *Nature* (left) and *Science* report the Bell Labs story.

in the physical sciences, however, remain blissfully ignorant. Some researchers interviewed for this article, for instance, were unaware of their own institution's misconduct policy. And some physics research centres — including Bell Labs — don't even have formal policies.

The prevailing view in the physical sciences seems to be that misconduct is someone else's problem. The American Physical Society's 'statement of integrity in physics' begins: "The physics community has traditionally enjoyed a well-deserved reputation for the maintenance of high ethical standards ... Indeed, the American Physical Society is one of the few professional societies which has not felt the need for a formal code of ethics."

One reason why researchers in the physical sciences feel cocooned from dishonesty is that their work often involves precise measurements that are relatively quick and easy to replicate once an experiment has been set up. This means that anyone who fabricates data is fairly likely to be found out. "If my graduate student or postdoc does something remarkable, I want to see the data, and usually, I would need to see the data two or three times done different ways," says Paul Canfield, a materials scientist at a University of Iowa. "That's a luxury I have because we're doing simple measurements."

In 'big science' areas of physics, where hundreds of individuals may be involved in a single experiment, the scope for fraud is even more limited. "I have to say that mis-

conduct of the kind alleged at Bell Labs is something that I have never heard of in high-energy physics,” says David Hitlin, who works on the BaBar experiment at the Stanford Linear Accelerator Center in California, which includes more than 500 physicists. At BaBar, says Hitlin, it would be extremely difficult to falsify results because several groups frequently analyse the same data in parallel.

Given these advantages, some researchers question the need for procedures to address misconduct. “You can sit around and have many meetings about how you can take care of rare cases if they ever occur, but is that the best use of your time?” asks Ronald Moon, who manages materials technology research for the electronics firm Agilent Technologies in Palo Alto, California. George Crabtree, a physicist who oversees materials science research at the Argonne National Laboratory in Illinois, agrees: “If you don’t need elaborate procedures, I don’t see a great value in putting them in place.” Indeed, Argonne has no internal policy on misconduct.

Eyes wide shut

To Nicholas Steneck, a historian of science at the University of Michigan who studies scientific misconduct, such statements reveal complacency. “Twenty years ago, people in the biomedical sciences said: ‘It’s not all that common, and it’s caught when it happens,’” says Steneck. “Well, 20 years of experience has shown that’s not the case.” No systematic studies have documented the extent of misconduct in the physical sciences, but Steneck suspects that many cases go unreported because the community lacks awareness.

Just as worrying, he fears that low awareness could also mean that the few researchers who are accused of misconduct might not receive a fair hearing. “People tend to over-react,” Steneck suggests. “They fire the person, make sure they don’t get a job anywhere, retract all the articles, and then say: ‘Whoops, wait a minute, do we need to do an investigation?’ Due process is still important.”



Readily checked, precise measurements help to safeguard the physical sciences, says Paul Canfield.

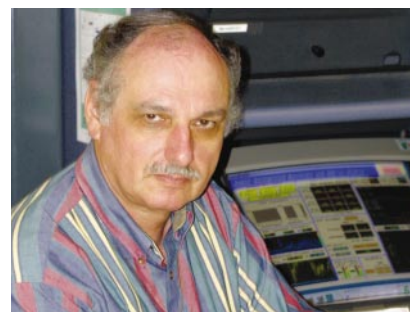
Researchers in the physical sciences could do a number of simple things to improve the situation, experts say. The first is simply to keep in mind that misconduct happens. “If you’re not prepared to look for it, you’re not likely to see it,” says Steneck.

Second, be aware of circumstances under which it is likely to occur. “Junior scientists are somewhat more likely to commit research misconduct,” says Helen Kerch, a programme manager in condensed-matter physics at the US Department of Energy (DOE). “Mentoring and guidance can really be beneficial.”

But how should you approach the situation if you suspect a protégé of foul play? One physicist who had to investigate a student offers this advice: “If you’re really feeling a little nervous about the research — and I think every scientist will have that feeling periodically — you sit down with the worker in your laboratory and say: ‘OK, I would like you to show me all the details, because I’ll be giving talks on this.’”

Finally, if you do have serious suspicions about a colleague’s work, says Kerch, remember to check your institution’s policies before doing anything else. “Many universities have a person who is basically in charge of misconduct,” she says, and that person should be consulted before any action is taken.

The agencies that fund research in the physical sciences also have some work to do, say experts in misconduct. The DOE,



David Hitlin: so many people work together in ‘big science’ that the scope for fraud is limited.

for instance, which is the largest provider of US federal funding for physics, is still working to implement the federal misconduct guidelines.

Forewarned is forearmed

Steneck stresses the importance of training. Since 1989, biomedical researchers have had to receive basic instruction in research integrity to be eligible for NIH training grants. But no such requirement applies to other classes of NIH grant, nor to grants awarded by the NSF and DOE.

Professional organizations could also do more to raise awareness, say experts. The ORI, for instance, sponsors joint conferences on research integrity with groups such as the Association of American Medical Colleges. “But it still hasn’t come into the physical sciences,” says Steneck.

So will the Schön case provide the jolt to make researchers in the physical sciences, and the organizations that represent them, address the issue? So far, there are no firm signs that this is happening. “The American Physical Society has no specific plans to revise its misconduct policies at this time, but is always alert to making changes in the future,” says a spokesperson. The American Chemical Society set up a committee to develop a policy on misconduct in April 2000, but it is still wrestling with whether simply to try to inform members about the issues, or to take sanctions against those who are found guilty of misconduct. “Quite frankly, I don’t think we’re making a whole lot of headway,” one committee member told *Nature*.

A good first step, suggest some experts, would be for agencies, societies or researchers to gather some baseline information on the incidence of misconduct in the physical sciences. “I don’t know how serious the problems are, but I do know that when they talk about having fewer problems they’re doing it on faith rather than evidence,” says Kerch. “I think it’s about time they took a look.” ■

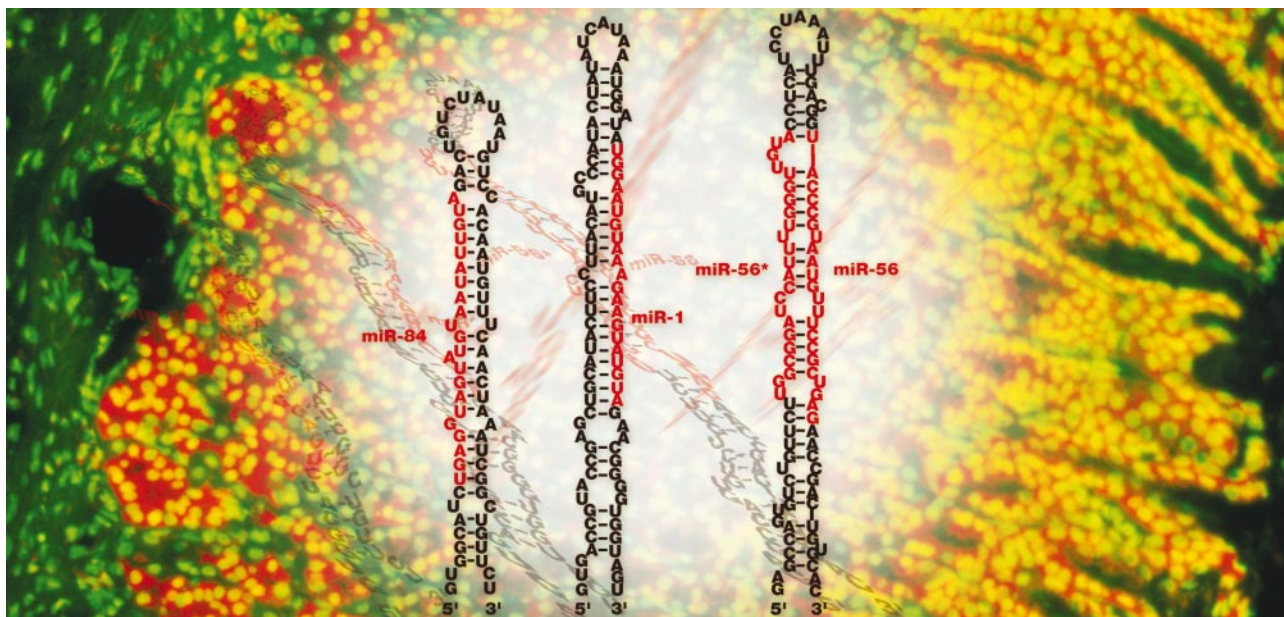
Geoff Brumfiel is *Nature’s* Washington physical sciences correspondent.

US federal guidelines

► www.ostp.gov/Science/html/sci_proj.html#Research%20Misconduct



See no evil: Nicholas Steneck believes physicists’ attitudes to misconduct betray complacency.



The brave new world of RNA

Most of the RNA transcribed from your genome doesn't make protein. Carina Dennis talks to the revolutionaries who believe that it functions in gene-regulatory networks that underlie the complexity of higher organisms.

Biology's 'central dogma', laid down in the 1950s, states that genetic information flows from DNA to RNA to protein. Since then, numerous studies have shown that RNA does more than simply serve this intermediary function. But is it time to cast dogma aside and completely rethink the role of RNA? The answer is yes,

according to a handful of geneticists. In multicellular organisms, they argue, the majority of RNA molecules are the principal actors in largely unexplored networks of gene regulation.

What's more, these researchers suggest that understanding RNA-based gene-regulatory networks may provide the key to explaining the difference between a yeast cell and a fruitfly, and between a fruitfly and you or me. "Complexity is hidden in the non-coding output of the genome," asserts John Mattick of the University of Queensland in Brisbane, Australia, who has published a series of papers that form an expansive but still speculative theoretical backdrop to the emerging view of RNA as a regulator of gene activity¹⁻³.

Many experts remain unconvinced. But as research in genomics reveals a vast array of RNAs that do not code for proteins, many biologists are finding it difficult to believe that they have no function. And with the discovery of a new class of non-coding RNAs, dubbed microRNAs (miRNAs), that seem to regulate the production of proteins from other genes, Mattick's ideas don't seem quite so outlandish.

In bacteria, most DNA codes for proteins. Geneticists have long known that the genomes of higher organisms are littered with non-coding DNA, much of which is

never read at all. But some of it is transcribed to RNA — your genes, for instance, contain non-coding 'introns' between the 'exons' that contain the instructions for making proteins. When a messenger RNA (mRNA) is transcribed from a gene, the introns are cut out and the exons spliced together. Many sequences outside protein-coding genes are also transcribed into RNA.

Mattick claims that introns and other non-coding RNAs may account for up to 98% of the transcriptional output of the human genome². He argues that non-coding RNAs interact with one another, with mRNAs, with DNA and with proteins to form networks that can regulate gene activity with almost infinite potential complexity.

It's an appealing idea, because comparisons of gene numbers don't seem to explain the difference between simple and complex organisms. We have only about two or three times as many protein-coding genes as the nematode *Caenorhabditis elegans* or the fruitfly *Drosophila melanogaster*, which, in turn, have only about twice as many as the yeast *Saccharomyces cerevisiae*. Higher organisms seem to bolster their complexity by mixing and matching protein domains to generate new combinations. But other plays may be needed to account for the vast complexity of humans and other vertebrates.



Denying dogma: John Mattick believes that non-coding RNAs underpin complex gene regulation.

Mattick likes to explain the workings of his proposed RNA-based networks in terms of computer science. The traditional biological view, based on the central dogma, is like a computer that would have to be rewired to perform each new calculation. An RNA network that interacts with all levels of the path from DNA to protein, however, could allow many different outputs to be derived from the same basic genetic circuitry, just as a computer's controlling software allows its processor to be easily reconfigured for a new task by changing the control codes. "It is not surprising that, for the evolution of complex systems, changes came about primarily in the control architecture rather than the components," Mattick argues.

Transcript trawl

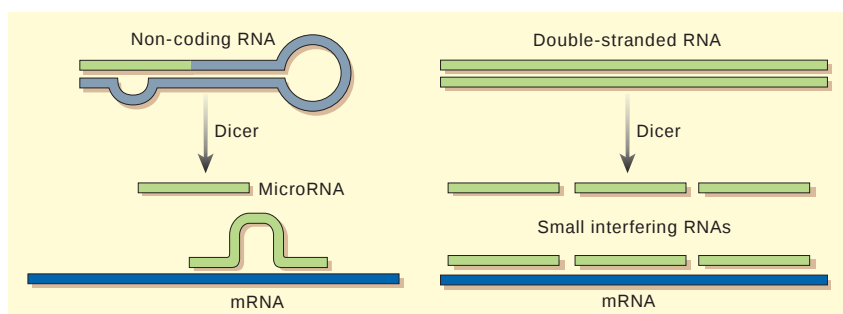
Accumulating the evidence to confirm or disprove this theory may take years. But analyses of the human genome are already showing that the realm of non-coding RNA is much bigger than most biologists had realized. In May, a team led by Thomas Gingeras, vice president for biological research at Affymetrix in Santa Clara, California, used DNA microarrays to probe each stretch of sequence across human chromosomes 21 and 22 for evidence of transcription. The researchers found that much more DNA was being transcribed than expected, given the number of genes that are thought to be present on these chromosomes⁴. "We saw ten times more sequence being transcribed than was predicted," says Gingeras.

Just showing that non-coding RNAs are common isn't enough to prove that most are involved in networks of gene regulation. But the discovery of miRNAs has given some credence to the idea. At roughly 22 nucleotides long, miRNAs passed unnoticed for a long time. The first genes coding for miRNAs — called *lin-4* and *let-7* — were identified in *C. elegans* in 1993 and 2000, respectively⁵⁻⁷. The miRNAs are cut from longer, hairpin-shaped RNAs transcribed from *lin-4* and *let-7*, and bind to specific target mRNAs, blocking their translation to proteins.

This new class of RNAs was initially considered to be peculiar to nematodes, until a team led by Gary Ruvkun of Harvard Medical School in Boston found that *let-7* is present in a diverse range of species, from vertebrates, including humans and zebrafish, to fruitflies and molluscs⁸.

The tally of miRNAs continues to increase. "It feels like a snowball about to turn into an avalanche," says Lawrence Hurst, an evolutionary geneticist at the University of Bath, UK. More than 150 miRNAs have now been identified from various animals, as a handful of groups have begun to prospect for them in earnest⁹⁻¹³. And this year has seen the first reports of their presence in plants^{14,15}.

Just how many miRNAs exist is not yet



The enzyme Dicer generates microRNAs that bind imperfectly to target mRNAs, blocking protein production (left); in RNAi, the RNAs made by Dicer bind tightly to mRNAs, targeting them for doom.

known. "We are still developing the technology to identify miRNAs, so it is difficult to make an estimate," says Victor Ambros of Dartmouth Medical School in Hanover, New Hampshire, who led one of the teams that first identified *lin-4*.

At the University of Münster in Germany, for instance, researchers led by Alexander Hüttenhofer are studying libraries of expressed sequences from mouse brain tissue. Whereas researchers hunting for protein-coding genes typically look for mRNAs in such extracts using methods that recognize characteristic structures called 'poly-A' tails, Hüttenhofer's team abandons this approach and uses gel electrophoresis to separate out smaller RNAs. "People looking for genes isolated extracts enriched for mRNA and threw away the remainder. We are the garbage people of the human genome project," says Hüttenhofer. So far, his team's search through the trash has yielded hundreds of small, non-coding RNAs¹⁶ — at least some of which are likely to be miRNAs.

The way in which miRNAs are generated has revealed an intriguing link to a gene-silencing mechanism known as RNA interference (RNAi), which is thought to defend cells from viruses and 'jumping genes'¹⁷. It springs

into action when the cell detects an unusual RNA with paired strands. An enzyme called Dicer cuts up the offending double-stranded RNA, creating fragments 21–25 nucleotides long, called small interfering RNAs (siRNAs). Single strands from these fragments bind to further copies of the original RNA, targeting them for destruction. RNAi can also be used experimentally to silence a cell's own genes, by adding double-stranded RNA sequences that match a gene's mRNA.

Making the cut

Studies of *lin-4* and *let-7* have revealed that Dicer cuts the miRNAs from the two genes' transcripts^{18,19}. They then bind to their target mRNA sequences, albeit less perfectly than the binding between siRNAs and their targets, causing them to bulge out (see diagram, above). The end result is also different: miRNAs merely prevent mRNAs from making proteins; they do not cause them to be destroyed. "My hunch is that RNAi is the ancestral pathway," says Ambros, who believes that aspects of its mechanism were subsequently co-opted to regulate the cell's own genes.

Because miRNAs do not destroy their targets, it is possible that they may impose a



Feline contrary: Sean Eddy contends that many non-coding RNAs are just "transcriptional slop".



Phillip Zamore says that it would be a cruel joke if conserved microRNAs have no function.

UMASS MED. SCHOOL



Small target: David Bartel says genetic screens have missed microRNAs because of their size.

temporary halt on the translation of mRNA to protein that can be relieved on cue. “Suppressing translation rather than destroying the transcripts seems a better way to modulate gene expression, as it leaves open the potential for switching translation back on again,” says Thomas Tuschl of the Max Planck Institute for Biophysical Chemistry in Göttingen, Germany.

The fact that the sequences of the miRNAs discovered so far are remarkably similar across different species suggests that most may share the gene-regulatory role played by *lin-4* and *let-7*. “It would be a cruel joke that nature has played on scientists if the sequences had been conserved in evolution but with no function,” says Phillip Zamore of the University of Massachusetts Medical School in Worcester.

Showing that most miRNAs have real biological significance, however, will require systematic attempts to see what happens if they are mutated — something that is easier said than done. “miRNAs have largely been missed by genetic screens because they are such small targets for mutagenesis,” says David Bartel of the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, who is now collaborating in a project to systematically create miRNA mutants in *C. elegans*.

Nevertheless, the limited evidence available supports the idea that miRNAs are involved in gene regulation. Some fruitfly mRNAs, for instance, contain sequence motifs that allow translation to the corresponding proteins to be shut down when required. And Eric Lai of the University of California, Berkeley, has found that some miRNAs have sequences that would bind perfectly to these motifs, suggesting that they may be switches that turn off protein production²⁰.

miRNAs are not the only non-coding RNA sequences that have been shown to play a role in regulating gene activity. A gene called *XIST*, for instance, encodes a large

RNA transcript that shuts down the ‘spare’ X chromosome in female mammals²¹. And in mice, a non-coding RNA dubbed *Air* has recently been shown to be involved in a gene-silencing mechanism that can shut down one of the two copies of a gene that are inherited, one from each parent²². Mattick is convinced that a wealth of further regulatory RNAs remains to be discovered. “This is just the tip of the iceberg,” he says.

But even among researchers studying non-coding RNAs, many believe that their gene-regulatory role could be relatively limited. Although many miRNAs may remain to be discovered, for now they are vastly outnumbered by the tens of thousands of known protein-coding genes. And as for the rest of the vast, uncharted landscape of non-coding RNA, it could still largely be garbage, after all.

Messy business

Many researchers believe that the process of transcribing RNA from DNA is inherently messy. “My opinion is simply that transcription might be a noisy process, and that a lot of RNAs are made for no good reason,” says Jean-Michel Claverie of the Institute of Structural Biology and Microbiology in Marseille, part of the CNRS, France’s national research agency. Sean Eddy, a computational biologist at Washington University in St Louis, Missouri, suspects that the cell rationalizes the costs of running a watertight operation against allowing some leakage — and leaky transcription wins out. “It’s cheap to make transcripts,” he argues. “The cell can tolerate a high level of transcriptional slop.”

But Hurst disagrees. “Many developmental pathways require exquisite control of gene expression,” he observes. “There are many reasons why a leaky genome is not desired.” And that view receives support from unpublished studies by Gingeras, who has compared the DNA sequences transcribed to non-coding RNA from human chromosomes 21 and 22 to the corresponding regions of the mouse and zebrafish genomes. The sequences are very similar, suggesting that they have been conserved by evolution because they perform some useful function.

Although the theory of RNA-based gene regulation seems attractive, putting it to the test won’t be easy. Building on Gingeras’s observations, one idea is to begin by looking for non-coding sequences that are conserved between the genomes of related species and

determining if they are transcribed into RNA, before attempting the painstaking bench work needed to investigate the transcripts’ function.

Eddy and others are also trying to devise computational strategies that can identify genes for functional non-coding RNAs from genome sequence data. Many of the non-coding RNAs that have so far been shown to have definite functions fold up into characteristic ‘secondary’ structures — so being able to predict these structures from DNA sequence data may help in the search for new non-coding RNAs.

Mattick, meanwhile, is looking for sequences within introns in yeast that match those in other protein-coding genes, reasoning that the former may produce RNAs that interact with the latter. In parallel, his team is conducting computer simulations of hypothetical RNA-based gene-regulatory networks, to see whether it is possible to generate immense complexity from simple beginnings simply by altering the regulatory network.

Potentially, gene regulation by non-coding RNA could not only provide the solution to the mystery of how higher organisms have generated such mind-boggling complexity from a relatively limited genetic palette, but may also explain much of the variation between individuals within a population — including the occurrence of some diseases. Intriguingly, Tuschl’s team has mapped the genetic aberration that causes chronic lymphoid leukaemia to a region of the genome that includes a gene encoding an miRNA, making it a plausible candidate.

But even if non-coding RNA turns out not to have overriding biological significance, the genome project’s garbage people have already sifted some gems from the trash. “Five or six years ago, people said we were wasting our time,” says Hüttenhofer. Today no one regards people studying non-coding RNA as time-wasters.

Carina Dennis is Nature’s Australasian correspondent.

1. Mattick, J. S. *Curr. Opin. Genet. Dev.* **4**, 823–831 (1994).
2. Mattick, J. S. & Gagen, M. J. *Mol. Biol. Evol.* **18**, 1611–1630 (2001).
3. Mattick, J. S. *EMBO Rep.* **2**, 986–991 (2001).
4. Kapranov, P. et al. *Science* **296**, 916–919 (2002).
5. Lee, R. C., Feinbaum, R. L. & Ambros, V. *Cell* **75**, 843–854 (1993).
6. Wightman, B., Ha, I. & Ruvkun, G. *Cell* **75**, 855–862 (1993).
7. Reinhart, B. J. et al. *Nature* **403**, 901–906 (2000).
8. Pasquinelli, A. E. et al. *Nature* **408**, 86–89 (2000).
9. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. *Science* **294**, 853–858 (2001).
10. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. *Science* **294**, 858–862 (2001).
11. Lee, R. C. & Ambros, V. *Science* **294**, 862–864 (2001).
12. Lagos-Quintana, M. et al. *Curr. Biol.* **12**, 735–739 (2002).
13. Mourelatos, Z. et al. *Genes Dev.* **16**, 720–728 (2002).
14. Llave, C. et al. *Plant Cell* **14**, 1–15 (2002).
15. Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B. & Bartel, D. P. *Genes Dev.* **16**, 1616–1626 (2002).
16. Hüttenhofer, A. et al. *EMBO J.* **20**, 2943–2953 (2001).
17. Hannon, G. J. *Nature* **418**, 32–39 (2002).
18. Grishok, A. et al. *Cell* **106**, 23–24 (2001).
19. Hutvagner, G. et al. *Science* **293**, 834–838 (2001).
20. Lai, E. C. *Nature Genet.* **30**, 363–364 (2002).
21. Avner, P. & Heard, E. *Nature Rev. Genet.* **2**, 59–67 (2001).
22. Sleuths, F., Zwart, R. & Barlow, D. P. *Nature* **415**, 810–813 (2002).



Interoperability calls for an unusual mix of skills

Computer-science expertise will be a key ingredient of a successful 'bioinformatics nation'.

Sir — Lincoln Stein, in his Commentary (*Nature* **417**, 119–120; 2002), argues that a web-services technology and a few agreed guidelines could help organize “a bioinformatics nation”. But can one technology be the ultimate solution for interoperability problems? I think it unlikely.

Although web services can be successfully leveraged for bioinformatics, similar claims were made not so long ago for CORBA (common object request broker architecture; www.omg.org). Until XML and web services appeared on the scene, many believed — I among them — that CORBA would be the technology of choice to solve interoperability problems in bioinformatics. It was already available five years ago. Notable efforts, from both industry and academic institutions, have been invested to create standards for the life sciences (see, for example, www.omg.org/technology/documents/formal/omg_

[lsr_specs.htm](http://www.omg.org/technology/documents/formal/omg_lsr_specs.htm)). Nevertheless, CORBA was insufficient to solve the interoperability problems that continue to challenge bioinformaticians.

Interoperability among biological databases and tools is a complex problem, only remotely influenced by a given technology, or by the fact that bioinformaticians may not follow what Stein calls a “data provider’s code of conduct”. For instance, I believe that interoperability is also limited by discipline-related factors. Today’s bioinformaticians have been trained primarily in scientific backgrounds (biology, chemistry, physics, mathematics and so on) that prepared them well for many topics in bioinformatics, but left them lacking sufficient knowledge of computer-science methods. The result is that bioinformaticians sometime ignore elementary design principles routinely applied by

software engineers. The lack of interoperability among the tools, and the databases built by this community, to some extent reflect the heterogeneity of the backgrounds in the community itself. I think that reversing this tendency will involve more than the steps suggested by Stein.

In particular, significant investments will be required to provide students with a multidisciplinary education that adds computer science and software engineering to their curriculum. In the short term, multidisciplinary teams can be assembled that bring together scientists and engineers. A cultural clash can be expected, but isn’t this the main challenge of interoperability?

Fabien Campagne

Institute for Computational Biomedicine, Mount Sinai School of Medicine, 1 Gustave L. Levy Place, Box 1218, New York, New York 10029, USA

Rebel army need not be a barrier to conservation

Sir — We welcome your highlighting of conservation problems in the Visayan region in the News story “Conservationists under fire in the Philippines” (*Nature* **416**, 669; 2002) profiling the work of the Philippine Endemic Species Conservation Project (PESCP). But it is disappointing that the security issues raised were not given appropriate context and explanation.

Your article implied, on the strength of one incident, that Negros Island is unsafe for tropical-forest conservation work. This is not the case. The local opposition to the PESCP on Panay Island that you report may reflect differences in campaigning style, as this is not our experience for non-campaigning organizations on Negros Island. The armed group who confronted the PESCP is likely to belong to the Revolutionary Proletarian Army — Alex Boncayao Brigade (RPA-ABB), whose agenda is political, not conservation-orientated. The incident probably arose from a lack of awareness by both parties.

Coral Cay Conservation has been working on Negros Island for seven years. It has deployed, without opposition, more than 2,000 volunteer field personnel to several areas including, for the past three years, the North Negros Forest Reserve (where the PESCP was located) as part of the community-based Negros Rainforest

Conservation Project (NRCP). Personnel often work from satellite camps in the forest and have met no hostility.

Security is an important concern, as the RPA-ABB has members based in and around the North Negros Forest Reserve. Their lack of hostility towards NRCP people probably reflects the fact that they know who we are and what we’re doing. The successful and unhindered efforts of the NRCP have been aided by communication at all levels. Coral Cay Conservation liaises closely with the British Embassy and the Foreign and Commonwealth Office, but it also works in partnership with a local group, the Negros Forests and Ecological Foundation, engaging all stakeholders (government, non-government and local communities) in the area, including the RPA-ABB.

Local stakeholders are keen to increase conservation activity on Negros Island. So, although it is important to promote the work of PESCP, it is also vital to address security issues in a local context.

Craig Turner, Pete Raines

Coral Cay Conservation, The Tower, 125 High Street, Colliers Wood, London SW19 2JG, UK

descendants have decided that the descendants of his slave Sally Hemings cannot join their family club, despite DNA evidence published in *Nature* suggesting that he fathered at least one of her children. In fact, that evidence suggests no specific Jefferson. The story also asserts that insistence on conclusive proof that the father was Thomas “cuts no ice” in academia. In fact, 13 noted scholars from 12 universities reviewed the historical and DNA evidence and unanimously found the paternity “by no means proven”². This ‘Scholars Commission’ — convened by the Thomas Jefferson Heritage Society but acting independently — criticized *Nature* for fostering misunderstanding of DNA evidence in the wider historical debate.

Steven T. Corneliusen

Thomas Jefferson National Accelerator Facility, 12000 Jefferson Avenue, Newport News, 23606 Virginia, USA

1. Check, E. *Nature* **417**, 213 (2002).
2. *The Jefferson-Hemings Scholars Commission Report on the Jefferson-Hemings Matter*, 12 April 2001, www.geocities.com/tjshcommission (to be published by Carolina Academic Press).

Jury out on Jefferson’s alleged descendants

Sir — Your News report¹ “Jefferson’s descendants continue to deny slave link” states that Thomas Jefferson’s

Correction There was an error in the first sentence of ref. 1. The word ‘ancestors’ was mistakenly used instead of ‘descendants’. The sentence should have read: “Thomas Jefferson’s descendants have decided that the descendants of one of his slaves cannot join their family club...”

Networking comes of age

How an old concept has grown new wings.

Linked: The New Science of Networks

by Albert-László Barabási

Perseus Publishing: 2002. 280 pp. \$26

Nexus: Small Worlds and the Groundbreaking Science of Networks/Small World: Uncovering Nature's Hidden Networks

by Mark Buchanan

W. W. Norton/Weidenfeld & Nicolson: 2002. 235 pp. \$25.95/£18.99

Sidney Redner

The past few years have seen an explosion of interest in so-called complex networks. In the physical sciences we are familiar with regular lattice networks of atoms, in which the local environment of each atom is identical. Such systems have traditionally been used to model the cooperative behaviour of interacting lattice systems — such as phase transitions — in which interactions are mediated by the underlying lattice.

However, recent work by Steven Strogatz and Duncan Watts on small worlds and by Albert-László Barabási and Réka Albert on scale-free networks has suddenly enlarged our notions of what actually constitutes a network. For example, in social settings it is clear that the acquaintance network formed by a collection of individuals is strongly heterogeneous. Some people are essentially reclusive and have few links to the outside world, whereas others are linked to a wide circle of friends. It would be inappropriate to describe such friendship networks as a regular lattice; something more akin to an airline route map, with a large number of poorly linked nodes and a few well-linked major 'hubs', seems more relevant. In fact, many man-made and natural networks, with the Internet and the web being two of the most obvious examples of the former type, appear to have just this kind of interconnection. Cooperative behaviour in such networks, as in the transmission of viruses among people or computers, is drastically different from that in lattices, and has many important epidemiological implications.

In *Linked*, Barabási presents an entertaining introduction to this vital field at a level that is generally accessible to the layperson interested in modern science. Barabási has made seminal contributions to the characterization of complex networks and has become an authoritative and forceful spokesman for the field. His enthusiasm is apparent throughout the narrative, a feature that makes for good reading. The book unravels many of the intriguing features of complex networks and will greatly enlarge a



Order in disguise: the pattern of streams in the Mississippi Delta follows a power-law distribution.

layperson's conceptions of what networks are all about.

The early sections of the book are interesting and informative. They include, for example, the stories about the original work on the six degrees of separation (the idea that just five links can connect any two people), the many amusing tales about the legendary mathematician Paul Erdős and his contributions to random graph theory, and the advances on small-world networks by Strogatz and Watts. Many applications are introduced with compelling examples, such as the AIDS epidemic to describe virus propagation through complex networks, or the early development of the Internet. It is very easy to be drawn in; many sections of the book have the feel of entertaining storytelling during conversations at a pavement café.

The middle chapters — called links — present a personalized account of recent advances in the field, including the many

substantial contributions made by the author and his research group. Among these are the original formulation of the scale-free network and its myriad of applications, the 'fitness' model, the behaviour of networks under attack, the spread of viruses and the role of complex networks in living organisms. Again, much of the work is presented in a highly digestible form (the numbered subsections help considerably here) and will pique the interest of many readers.

Although I am enthusiastic about the book, I do have a few minor quibbles. The link describing Einstein's legacy seems mis-cast. Not enough background has been given for a layperson to appreciate Bose–Einstein condensation and the analogy to the fitness model. This section also contains some technical flaws. There are also a few instances of unexplained jargon, such as: "most quantities in nature follow a bell curve", "the connectivity distribution of the Internet

routers follows a power law”, the notion of NP complete problems and the travelling-salesman problem. Regrettably, the basic Figure 6.1, which is meant to illustrate a power-law degree distribution, is of poor quality and it would be more informative to plot power-law behaviour on a double-logarithmic scale. Overall, however, Barabási has pitched the presentation at a highly accessible level.

Nexus is a very similar book. Written by Mark Buchanan, a science writer and physics doctorate, it gives a cogent and engaging description of recent developments in complex networks. There is much overlap with *Linked*, in both content and style. But Barabási's book is more focused and follows many of his own very important contributions, whereas Buchanan's provides a slightly broader perspective but sometimes strays from the topic of networks.

Both books are extremely valuable contributions to the popular-science literature and I enjoyed them immensely. I would enthusiastically recommend them to my own family members. It is remarkable and gratifying to see excellent popularizations of a burgeoning field of scientific research while many developments are still in progress. ■

Sidney Redner is in the Center for BioDynamics, the Center for Polymer Studies, and in the Department of Physics, Boston University, Boston, Massachusetts 02215, USA.

A right royal feud

The Correspondence of John Flamsteed, First Astronomer Royal: Volume 3, 1703–1719
Compiled and edited by Eric G. Forbes, Lesley Murdin & Frances Willmoth
Institute of Physics: 2001. 1,104 pp.
£199, \$298.50

Owen Gingerich

Had this long-awaited third and final volume of John Flamsteed's correspondence appeared 170 years ago, it would have created consternation, incredulity and a serious scandal in Victorian Britain. Sir Isaac Newton, a national hero, the paragon of scientific insight and moral rectitude, apotheosized in Westminster Abbey, would have been exposed as a conniving, arrogant bully who had embarked on a campaign of defamation against the upright, albeit straitlaced, mathematicus regius. This book is the documentation of Newton's confrontation with the hapless Reverend John Flamsteed, the first astronomer royal.

Today the shock value is considerably diminished, for the news was in fact broken in 1835 to a disbelieving British public by Francis Baily in his *An Account of the Rev. John Flamsteed, the First Astronomer-Royal*. But



Although Flamsteed (above) showed exceptional ability, in Newton's eyes he was a mere “puppy”.

now, at long last, the entire correspondence is available, with detailed notes to clarify numerous references blurred by the passage of nearly three centuries. The previous two volumes showed the once-friendly collaboration between Newton and Flamsteed gradually fade to an arm's-length relationship. In this, the thickest of the three volumes, the interaction comes to a boiling feud, climaxing in a letter dated 22 December 1711 when Flamsteed informed his former assistant and confidant, Abraham Sharp, that a furious Newton had called him (horrors!) a puppy.

In hindsight we can see that Flamsteed was an extraordinarily competent, obstinately self-motivated and highly driven observational astronomer worthy of having his correspondence edited in detail. But he was not a genius like Newton, who, in his arrogance, recognized his superiority and could scarcely disguise his feelings that in comparison the astronomer royal was a mere pup. Flamsteed was not about to concede his lesser ability, and a substrate of jealousy and attempted one-upmanship is scarcely concealed in his letters to Sharp, who provided the perfect foil. (The correspondence between this pair comprises nearly half of the surviving letters in this volume, as opposed to those whose existence is merely deduced.)

In the early years covered in this volume, the relationship between these two egos was still sufficiently cordial for Newton to send Flamsteed a copy of his *Opticks* in 1704. “I was delighted with it till I came to pag 72.73 where I found he committed a great fault,” Flamsteed confided to James Pound, another astronomer. “Mr N came down to see me, I shewed him the error of his assertion and an easy experiment whereby it might be proved

on our earth that the rays of light spreading had no such monstrous effects as he imagined. He gave me the hearing quietly and made me no answer.”

But from there on the relations went downhill. Newton “has been lately much talkt of but not much to his advantage”, Flamsteed wrote to Sharp in 1709. “Our [Royal] Society is ruined by his close politick and cunning forecast [design] I fear past retrieveing for our Drs Transactions have been twice burlesqt publicly and now we have had none published I think this four moneths.”

By the time the second edition of Newton's *Principia* was published in 1713, communications between the two men had completely deteriorated. Flamsteed zeroed in on the lunar theory — precisely the area where Newton had been so keen to get the astronomer royal's observations, and the root of the great controversy between them. Flamsteed wrote to Sharp that Newton's sixth equation “is not allowed by the Heavens”, implicitly admitting that he had 112 observations that Newton had not seen. The three-body lunar problem was very complex, and it would be many years before the gravitational theory was tested using enough terms to give predictions independently of empirically determined coefficients. If Newton had had access to Flamsteed's observations, no doubt he would have fudged the theory enough to bring the coefficients into line with the empirical data.

Meanwhile, Flamsteed gave grudging admiration to the gravitational theory, although he probably didn't understand it very well. It was becoming increasingly obvious that the periods of Jupiter and Saturn had long-term variations, and that these might be explained by their mutual attractions. About the great comet of 1680, Flamsteed wrote: “This is the consequence of Keplers doctrine of Magneticall fibres, improved by Sir Chr. Wren and prosecuted by Sir I. Newton and I thinke I can lay some claime to a part of it...” Newton had earlier argued that two comets were involved, whereas Flamsteed believed it was a single comet that had undergone a hairpin turn near the Sun. When Newton capitulated, it produced strong input for his concept of universal gravitation. Flamsteed, however, had placed the turn in front of the Sun, rather than around it, and it was just his fantasy that he deserved part of the credit for the gravitational theory.

Although the relations with Newton take centre stage in this volume, there are many fascinating subthemes as well. In 1714 Flamsteed was appointed a judge for proposals seeking the £20,000 prize for a method to find longitude at sea, and hopeful inventors put their cases to him. Detailed descriptions of brilliant auroral displays fill a number of letters. Their cause was still unknown, but

Flamsteed was intrigued, and near the end of his life he remarked to Sharp: "You see that neither Age nor Infirmitys, nor any discouragements hinder me from seeking after knowledge and Truth."

Particularly revealing were the now-forgotten amateur astronomers who wrote detailed letters comparing their observations with a variety of competing astronomical tables. Clearly the "great inequality" of Jupiter and Saturn was raising its head, not to be solved till the end of the eighteenth century. In the final paragraph Flamsteed wrote to Sharp, a month before his death, he remarked: "but Saturn and Jupiter will find worke for those that come after us." ■

Owen Gingerich is at the Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA.

Evolution by design?

No Free Lunch: Why Specified Complexity Cannot Be Purchased without Intelligence

by William A. Dembski

Rowman & Littlefield: 2002. 432 pp. \$35, £27

Brian Charlesworth

Newton believed that his discoveries revealed the Universe to be subject to the "counsel and dominion of an intelligent and powerful Being". A century later, Laplace famously remarked that "I had no need of that hypothesis", after completing his revision of celestial mechanics. The whole enterprise of modern science is built on the assumption that nature can be understood without appealing to the intervention of gods or goblins. Most people would agree that it has been remarkably successful.

William Dembski and his fellow advocates of "Intelligent Design" want to turn the clock back. They claim that darwinian evolution is inadequate to explain the intricate adaptations that can be seen in even the simplest single-celled organism. Their position differs from that of the biblical creationists in that they accept the scientific estimates of the age of the Earth and the Universe, and even allow a limited role for evolutionary processes such as natural selection. Their arguments are dressed up with a good deal of philosophical and mathematical formalism, but conclude with an appeal to the continual intervention of an unobservable designing intelligence in the course of nature. This smacks of the Middle Ages.

Dembski uses two arguments, neither of which is new. The first is based on the problem of "specified complexity": the improbability of assembling a functioning complex

system by randomly combining numerous individual components. This is a variant of Hoyle and Wickramasinghe's old claim that the probability of assembling a functioning protein out of a random set of amino acids is similar to that of Concorde being constructed by a tornado hitting a junkyard. As Darwin made abundantly clear, this is not the way that evolution by natural selection is thought to work. Rather, a step-by-step adjustment of individual characters occurs, each of which is advantageous in terms of darwinian fitness.

After a good deal of twisting and turning, Dembski concedes that this model of adaptive evolution can indeed work, and produce the appearance of design. But he then argues that there is still no "free lunch", since we have to be able to explain the trajectory over time of the relationships between character states and fitness. According to him, this requires the intervention of an intelligent designer, rather than the interaction of an evolving population with a changing physical and biotic environment. This ignores the large body of biological evidence on the emergence of evolutionary novelties in response to new environments, seen on a small scale in island radiations such as Darwin's finches, and on a large scale in such events as the evolution of mammalian groups after the extinction of the dinosaurs.



Rocky blues

Azurite, found in Chessy, France. Taken from *Photographic Guide to Minerals of the World* by Ole Johnsen (Oxford University Press, £17.99, \$29.95), originally published in Danish.

The second argument is that of "irreducible complexity" — that a piece of biological machinery made up of many integrated components, such as a bacterial flagellum, may fail to work if even one of them is removed. This is taken to imply that it could not have evolved via a series of functional intermediates. Again, this argument was considered by Darwin in his famous discussion of the eye. Darwin pointed out that, although we have no evidence about the actual historical course of events that led to the evolution of the vertebrate eye, we do have many different types of eyes represented among living animals, ranging from simple light-sensitive cells to the elaborate camera-like eyes of cephalopods and vertebrates. A series of gradual steps leading from the simplest up to the most elaborate eye can be reconstructed, each of which involves a light receptor that is of use to its possessor.

Like other historical sciences, evolutionary biology attempts to interpret the past in the light of processes that can be observed to be acting today. The direct record of what actually happened, as revealed in the fossil record, is necessarily incomplete, especially with respect to very remote events such as the origin of life itself. Nevertheless, we can ask if what we see in the living world, both in the past and today, is consistent with what is expected from our models of evolution. The scientific literature on evolution is full of examples of people trying to do just that. For example, if a lack of intermediates is a product of the incompleteness of the fossil record, we would expect to see more and more discoveries of intermediate forms as time goes by. Human evolution provides an excellent example of this, with the fossil discoveries of the twentieth century providing a resounding confirmation of Darwin's hypothesis of an African origin of modern humans.

In contrast, Dembski prefers to fill up the gaps in our knowledge by invoking an agent whose acts have no observable consequences. He smugly refuses to provide any details of what the designer has in mind. His theory can explain anything, and therefore explains nothing. If one tries to imagine what the designer is like by looking at the

facts of biology, one is likely to come to rather uncomfortable conclusions.

As J. B. S. Haldane wrote in 1932: "Blake expressed some doubt as to whether God had made the tiger. But the tiger is in many ways an admirable animal. We have now to ask if God made the tapeworm. And it is questionable whether an affirmative answer fits in either with what we know about the process of evolution or what many of us believe about the moral perfection of God." ■

Brian Charlesworth is at the Institute for Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK.

The bigger picture

Tamas Vicsek

If a concept is not well defined, it can be abused. This is particularly true of complexity, an inherently interdisciplinary concept that has penetrated a range of intellectual fields from physics to linguistics, but with no underlying, unified theory. Complexity has become a popular buzzword that is used in the hope of gaining attention or funding — institutes and research networks associated with complex systems grow like mushrooms.

Why and how did this vague notion become such a central motif in modern science? Is it only a fashion, a kind of sociological phenomenon, or is it a sign of a changing paradigm of our perception of the laws of nature and of the approaches required to understand them? Because almost every real system is inherently complicated, to say that a system is complex is almost an empty statement — couldn't an Institute for Complex Systems just as well be called an Institute for Almost Everything? Despite these valid concerns, the world is indeed made of many highly interconnected parts on many scales, the interactions of which result in a complex behaviour that requires separate interpretations of each level. This realization forces us to appreciate the fact that new features emerge as one moves from one scale to another, so it follows that the science of complexity is about revealing the principles that govern the ways in which these new properties appear.

In the past, mankind has learned to

understand reality through simplification and analysis. Some important simple systems are successful idealizations or primitive models of particular real situations — for example, a perfect sphere rolling down an absolutely smooth slope in a vacuum. This is the world of newtonian mechanics, and it ignores a huge number of other, simultaneously acting factors. Although it might sometimes not matter that details such as the motions of the billions of atoms dancing inside the sphere's material are ignored, in other cases reductionism may lead to incorrect conclusions. In complex systems, we accept that processes that occur simultaneously on different scales or levels are important, and the intricate behaviour of the whole system depends on its units in a non-trivial way. Here, the description of the entire system's behaviour requires a qualitatively new theory, because the laws that describe its behaviour are qualitatively different from those that govern its individual units.

Take, for example, turbulent flows and the brain. Clearly, these are very different systems, but they share a few remarkable features, including the impossibility of predicting the rich behaviour of the whole by merely extrapolating from the behaviour of its units. Who can tell, from studying a tiny drop or a single neuron, what laws describe the intricate flow patterns in turbulence or the patterns of electrical activity produced by the brain? Moreover, in both of these systems (and in many others), randomness and determinism are both relevant to the system's overall behaviour. Such systems exist on the edge of chaos — they may exhibit almost regular behaviour, but also can change dramatically and stochastically in time and/or space as a result of small changes in conditions. This seems to be a general property of systems that are capable of producing interesting (complex) behaviour.

Knowledge of the physics of elementary particles is therefore useless for interpreting behaviour on larger scales. Each new level or scale is characterized by new, emergent laws that govern it. When creating life, nature acknowledged the existence of these levels by spontaneously separating them into molecules, macromolecules, cells, organisms, species and societies. The big question is whether there is a unified theory for the ways in which elements of a system organize themselves to produce a behaviour that is typical of large classes of systems.

Interesting principles have been proposed in an attempt to provide such a unified theory. These include self-organization, simultaneous existence of many degrees of freedom, self-adaptation, rugged energy landscapes,

Complexity

The laws that describe the behaviour of a complex system are qualitatively different from those that govern its units.

and scaling (for example, power-law dependence) of the parameters and the underlying network of connections. Physicists are learning how to build relatively simple models that can produce complicated behaviour, whereas those who work on inherently very complex systems (such as biologists and economists) are uncovering ways to interpret their subjects in terms of interacting, well-defined units (such as proteins).

What we are witnessing in this context is a change of paradigm in attempts to understand our world as we realize that the laws of the whole cannot be deduced by digging deeper into the details. In a way, this change has been wrought by the development of instruments. Traditionally, improved microscopes or bigger telescopes are built to gain a better understanding of particular problems. But computers have allowed new ways of learning. By directly modelling a system made of many units, one can observe, manipulate and understand the behaviour of the whole system much better than before, as in the cases of networks of model neurons and virtual auctions by intelligent agents, for example. In this sense, a computer is a tool that improves not our sight (as does the microscope or telescope), but rather our insight into mechanisms within complex systems.

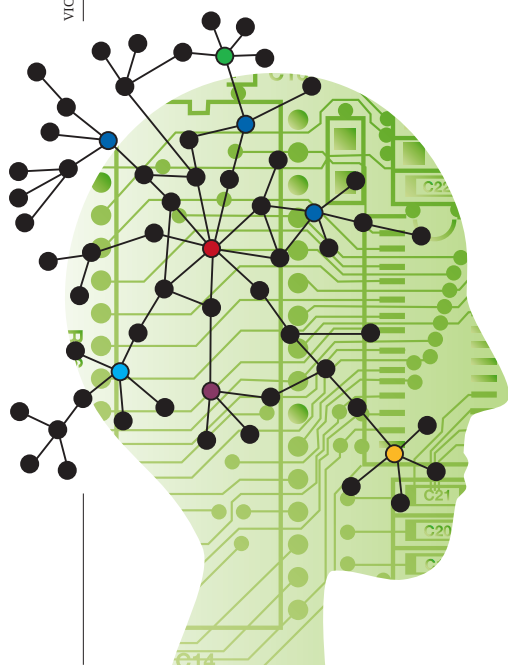
Many scientists implicitly assume that we understand a particular phenomenon if we have a (computer) model that provides results that are consistent with observations and that makes correct predictions. Yet such models make it possible to simulate systems that are far more complex than the simplest newtonian ones that allow deterministic, accurate predictions of future events. In contrast, models of complex systems frequently result in a new conceptual interpretation of the behaviour. The aim is to capture the principal laws behind the exciting variety of new phenomena that become apparent when the many units of a complex system interact. ■

Tamas Vicsek is in the Department of Biological Physics, Eötvös University, Budapest, Pázmány Stry 1A, H-1117 Hungary.

FURTHER READING

Waldrop, M. M. *Complexity* (Simon & Schuster, New York, 1993).
Gell-Mann, M. *Europhys. News* **33**, 17 (2002).
www.comdig.org
Nature Insight, Complex Systems *Nature* **410**, 241–284 (2001).

VICKY ASKEW



Hominid revelations from Chad

Bernard Wood

The story of human origins in Africa takes a twist with the description of a 6–7-million-year-old cranium from Chad. The discovery hints at the likely diversity of early hominids.

A single fossil can fundamentally change the way we reconstruct the tree of life. More than 75 years ago, Raymond Dart's description¹ of the Taung skull from southern Africa wrought such a transformation with regard to human evolution. Dart provided hard evidence to support Darwin's prediction that the roots of human evolutionary history run deepest in Africa.

A fossil cranium (Fig. 1), discovered by Michel Brunet and his colleagues and described in this issue, marks a similar turning point in our understanding of human origins. Discussion of the cranium and associated fossils is on page 145 (Brunet *et al.*²), with presentation of the contextual evidence (Vignaud *et al.*³) on page 152. The fossils — the cranium, a jaw fragment and several teeth — belong to a primitive human precursor, or hominid, that is an astonishing 6–7 million years old. The transformation wrought here is more nuanced than Dart's, but it is as fundamental. Here we have compelling evidence that our own origins are as complex and as difficult to trace as those of any other group of organisms.

For almost 150 years⁴ it has been suggested that modern humans are more closely related to the African apes than they are to the orang-utan. Nowadays, evidence from both bones and teeth^{5–7}, and soft tissues (muscles, nerves, and so on)⁸, and from molecular and DNA analyses^{9,10}, support the view that modern humans and chimpanzees are particularly closely linked. When the DNA differences are calibrated by using palaeontological evidence, they indicate that the hypothetical ancestor of modern humans and the chimpanzee lived between about 5 and 7 million years ago.

The hominid fossil record outside Africa has stubbornly failed to break the 2-million-year barrier. Thus, if the 'molecular clock' keeps reasonably good time, between 3 and 5 million years or so of our independent evolution took place on the African continent. Four regional 'windows' provide fossil evidence relevant to our early evolutionary history. The southern African window was revealed by Dart in 1925 when the first (and only) hominid fossil from Taung, near Kimberley, was recognized; since then, neighbouring cave sites have provided a rich fossil record that stretches back to around 3–3.5 million years ago¹¹. The East African window comprises sites along the Eastern, or Gregory, Rift Valley, from close to the Gulf of Aden

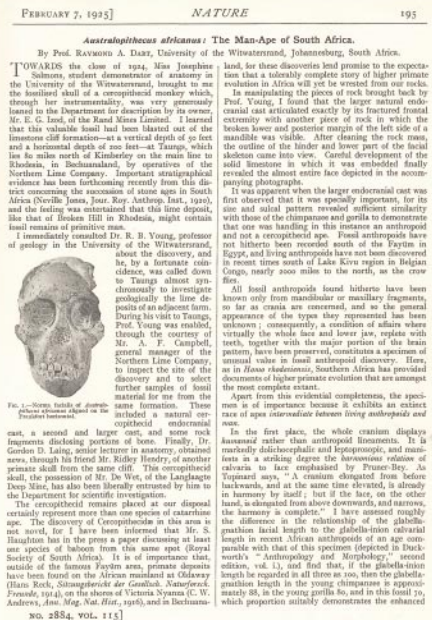


Figure 1 The cranium of the newly described *Sahelanthropus tchadensis* and (below) the opening page of Raymond Dart's 1925 description of the Taung skull.

in the north to northern Tanzania in the south. The sites are associated with sedimentary basins or the rivers that fed or drained them. Two of them, Middle Awash in Ethiopia^{12,13} and Lukeino in Kenya¹⁴, have so far provided the oldest evidence of creatures that are plausible human ancestors.

The two remaining regions, Malawi and Chad, were, until now, more like spy-holes than windows. Malawi has provided evidence of one of the large-toothed hominid species, probably *Paranthropus aethiopicus* (see Fig. 2, overleaf). The first 'early hominid' from Chad, *Tchadanthropus uxoris*, found in 1961, turned out to be the face of a modern human skull that had been so eroded by wind-blown sand that it mimicked the appearance of an australopithecine, a primitive type of hominid. The second Chad hominid, *Australopithecus bahrelghazali*, discovered¹⁵ at a site called Koro Toro in 1995, is an authentic australopithecine and alerted palaeontologists to the potential of central West Africa.

Four areas in and around the Chad basin have yielded mammalian fossils, but it is one locality, TM 266, in the oldest of these areas — Toros-Menalla in the Djurab Desert — that provided Brunet's team with the fossils they describe in this issue. The discovery is a tribute to the tenacity of Brunet, Vignaud and their scientific colleagues, and to their intrepid local field team. The sand-laden wind



blows incessantly and the fossil layers are difficult to detect: they are at most a few metres thick and a far cry from the banks of sediment that we are used to seeing in pictures of Olduvai Gorge in Tanzania, and the like. Yet, despite the harsh present-day environment, the vertebrate fossils are well preserved, and the hominid cranium (designated TM 266-01-060-1) is remarkably complete.

Absolute — isotope-based — dating methods cannot be applied to the fossil layers at Toros-Menalla because there are no ash layers to provide the necessary argon and

potassium. Nor are the sediments suitable for magnetism-based dating methods. Instead, the team matched the rich vertebrate fossil record at TM 266, consisting of examples of 44 different groups, with the equivalent record from sites in East Africa that have absolute dates. The best matches are with two sites in Kenya: the Lukeino Formation of the Tugen Hills (which dates to about 6 million years ago) and the Nawata Formation at Lothagam (5.3–7.4 million years). The upshot is a reliable age estimate of about 6–7 million years for the Toros-Menalla fossils.

The researchers compared their new evidence with what has been published about two other claimants for the title of 'earliest hominid': *Ardipithecus ramidus* from the Middle Awash^{12,13} and *Orrorin tugenensis* from Lukeino¹⁴. They satisfied themselves (and others, myself included) that the teeth of the new fossils are taxonomically distinctive, and accordingly assigned the fossils to a new species and genus, *Sahelanthropus tchadensis*.

What was the role of *S. tchadensis* in the evolution of chimpanzees and modern humans? The latter two look very different, but the differences between the earliest ancestors of chimpanzees and modern humans are likely to have been more subtle. The conventional presumption is that the human–chimp common ancestor, and the earliest members of the chimp lineage, or clade, would have been adapted for life in the trees, with the trunk held either horizontal or upright and with the forelimbs adapted for knuckle-walking on large branches or on the ground. This would have been combined with projecting faces that accommodated elongated jaws bearing relatively small chewing teeth and, in males, large upper canine teeth that would have worn against the lower premolars.

Early hominids at the base of our own clade, in contrast, would have been distinguished by at least some skeletal and other adaptations for an upright posture and bipedal walking and running, linked with a chewing apparatus that combined proportionally larger chewing teeth, modest-sized male canines that wore only at the tip of the crown, and some evidence of an increase in brain size. Against these criteria it is the face, jaw and canines of *S. tchadensis* that point to its being a hominid, at (or at least close to) the base of the modern human clade.

There are two current hypotheses about human origins and the early stages of hominid evolution. According to the linear, or 'tidy', model¹⁶, the distinctive hominid anatomy evolved only once, and was followed by a ladder-like ancestor–descendant series. In this model there is no branching (cladogenesis) until well after 3 million years ago. The bushy, or 'untidy', model sees hominid evolution as a series of successive adaptive radiations — evolutionary diversi-

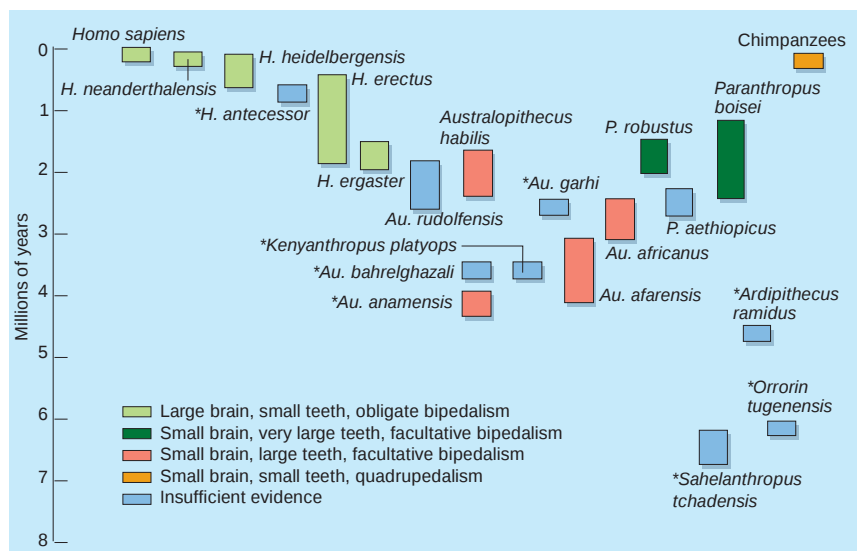


Figure 2 The known fossil record of hominids, including *S. tchadensis*^{2,3}, also showing ourselves (top left) and the chimpanzee (top right). Extinct species are indicated with the dates of the earliest and latest fossil evidence, but these are likely to increase and decrease, respectively, especially for the less well-known examples. Species are assigned to one of four categories, based on brain and cheek-tooth size, and inferred posture and locomotion (we are obligately bipedal; facultative bipedalism is the ability to walk or run on two legs, or as a quadruped, according to circumstances). A fifth category is for 'insufficient evidence'. The species marked with an asterisk were all unknown a decade or so ago, an indication of the paucity of evidence, until recently, of hominid evolution between 1 and 4 million years ago. This comparatively rich record contrasts with the earlier part of the hominid fossil record. There are likely to be many 'undiscovered' species in the fossil record between 7 and 4 million years ago, and in reconstructing the early stages of human evolution in particular the incompleteness of data should always be acknowledged.

fication in response to new or changed circumstances — in which anatomical features are 'mixed and matched' in ways that we are only beginning to comprehend^{17,18}. This model, to which I subscribe, predicts that because of the independent acquisition of similar shared characters (homoplasy), key hominid adaptations such as bipedalism, manual dexterity and a large brain are likely to have evolved more than once¹⁹. So the evidence of one, or even a few, of the presumed distinguishing features of hominids might not be enough to link a new species with later hominids, let alone to identify it as the direct ancestor of modern humans.

What is remarkable about the chimpanzee-sized cranium TM 266-01-060-1 discovered by Brunet *et al.* is its mosaic nature. Put simply, from the back it looks like a chimpanzee, whereas from the front it could pass for a 1.75-million-year-old advanced australopithecine. The hominid features involve the structure of the face, and the small, apically worn, canine crowns. Other hominid features are found in the base of the cranium and in the separate jaw fragment. If we accept these as sufficient evidence to classify *S. tchadensis* as a hominid at the base, or stem, of the modern human clade, then it plays havoc with the tidy model of human origins. Quite simply, a hominid of this age should only just be beginning to show signs of being a hominid. It certainly should not have the face of a

hominid less than one-third of its geological age. Also, if it is accepted as a stem hominid, under the tidy model the principle of parsimony dictates that all creatures with more primitive faces (and that is a very long list) would, perforce, have to be excluded from the ancestry of modern humans.

In contrast, the untidy model would predict that at 6–7 million years ago we are likely to find evidence of creatures with hitherto unknown combinations of hominid, chimp and even novel features. Moreover, because it acknowledges substantial amounts of homoplasy, the model would further predict that certain structures — such as substantial brow ridges (which *S. tchadensis* has, as is evident in Fig. 1) — are likely to be unreliable for reconstructing relationships because creatures can share features such as brow ridges without necessarily inheriting them from a common ancestor²⁰. *S. tchadensis* is a candidate for the stem hominid, but in my view it will be impossible to prove that it is.

My prediction is that *S. tchadensis* is just the tip of an iceberg of taxonomic diversity during hominid evolution 5–7 million years ago. Its potentially close relationship with our own, hominid, twig of the tree of life is surely important. More notably, however, I think it will prove to be telling evidence of the adaptive radiation of fossil ape-like creatures that included the common ancestor of modern humans and chimpanzees. The fauna of the

Burgess Shale in Canada, which samples a bewildering array of invertebrate groups some 500 million years ago, is a famous example of diversity at the base of an adaptive radiation. Does *S. tchadensis* belong to the African ape equivalent of the Burgess Shale? ■

Bernard Wood is in the Department of Anthropology, The George Washington University, 2110 G Street NW, Washington DC 20052, USA. e-mail: bwood@gwu.edu

1. Dart, R. A. *Nature* **115**, 195–199 (1925).
2. Brunet, M. et al. *Nature* **418**, 145–151 (2002).
3. Vignaud, P. et al. *Nature* **418**, 152–155 (2002).
4. Huxley, T. H. *Evidence as to Man's Place in Nature* (Williams & Norgate, London & Edinburgh, 1863).
5. Groves, C. P. in *Systematics of the Great Apes* (eds Swindler, D. R. & Erwin, J.) 187–217 (Liss, New York, 1986).
6. Groves, C. P. & Paterson, J. D. J. *Hum. Evol.* **20**, 167–183 (1991).
7. Shoshani, J., Groves, C. P., Simons, E. L. & Gunnell, G. F. *Mol. Phylogenet. Evol.* **5**, 101–153 (1996).

Planetary science

Out on the edge

William B. McKinnon

The Kuiper Belt, beyond Neptune, is the third great domain of the Solar System, and home to the Pluto–Charon binary. What are the prospects for exploration of these distant worlds?

The discovery of the Kuiper Belt in the far regions of the Solar System is one of the great achievements of the space age. In addition to the small planet Pluto and its large moon Charon, the belt contains about 100,000 worlds greater than 100 km in diameter, as well as a vast number of smaller, cometary bodies¹. Unlike the domains of the terrestrial planets (Mercury to Mars) or the gas giants (Jupiter to Neptune), the Kuiper Belt has never been explored by spacecraft — although one mission, 'New Horizons', has been competitively selected and is in its final design phase². In May, a workshop* was convened to review progress since the last major conference on Pluto–Charon³ and to imagine the results of the possible spacecraft encounter in 2015 or 2016.

In astronomy, there is no substitute for resolution and Pluto and Charon are, to put it mildly, poorly resolved from Earth. One clear signature, however, is Pluto's rotational lightcurve — the variation of the planet's apparent brightness with time. Pluto's lightcurve is quite pronounced, both in terms of brightness and spectral features, and implies at least three separate types of surface terrain (W. Grundy, Lowell Observatory, Arizona): a bright, nitrogen-ice-rich terrain containing dissolved methane and carbon monoxide; another bright, reddish terrain dominated by methane ice; and a third dark, volatile-depleted terrain betraying only the slightest hint of the broad infrared absorp-

8. Gibbs, S., Collard, M. & Wood, B. J. *Anat.* **200**, 3–49 (2002).
9. Gagneux, P. V. & Varki, A. *Mol. Phylogenet. Evol.* **18**, 2–13 (2002).
10. Wildman, D., Grossman, L. I. & Goodman, M. in *Probing Human Origins* (eds Goodman, M. & Moffat, A. S.) 1–10 (Am. Acad. Arts Sci., Cambridge, MA, 2002).
11. Partridge, T. C., Shaw, J., Heslop, D. & Clarke, R. J. *J. Quat. Sci.* **14**, 293–298 (1999).
12. White, T. D., Suwa, G. & Asfaw, B. *Nature* **371**, 306–312 (1994).
13. Haile-Selassie, Y. *Nature* **412**, 178–181 (2001).
14. Senut, B. et al. *C. R. Acad. Sci. Paris, Earth Planet. Sci.* **332**, 137–144 (2001).
15. Brunet, M. et al. *Nature* **378**, 273–275 (1995).
16. Asfaw, B. et al. *Science* **284**, 629–635 (1999).
17. Wood, B. R. & Richmond, B. G. *J. Anat.* **197**, 19–60 (2000).
18. Leakey, M. G. et al. *Nature* **410**, 433–440 (2001).
19. Collard, M. C. & Wood, B. A. *Proc. Natl Acad. Sci. USA* **97**, 5003–5006 (2000).
20. Lieberman, D. E. in *Development, Growth and Evolution: Implications for the Study of the Hominid Skeleton* (eds O'Higgins, P. & Cohn, M. J.) 85–122 (Academic, London, 2000).

tions of water ice^{4,5}. Such a complex, variegated surface goes a long way towards explaining the peculiarities of Pluto's heat signature: the

planet simultaneously exhibits a nitrogen-dominated atmosphere in vapour-pressure equilibrium with nitrogen-frosted terrain at a temperature of 40 K, and warmer regions where volatile ices have burned off⁶. In this regard, Pluto is Mars-like in its surface–atmosphere interaction.

The lightcurve data only hint at the complexity of Pluto's surface. A higher-resolution map — derived from the mutual eclipses and transits of Pluto and its moon in the 1980s — shows that even within the dark, volatile-depleted regions there exist significant visual colour differences⁷, although these differences are not as extreme as those seen in the Kuiper Belt population as a whole (S. Tegler, Northern Arizona Univ.).

By contrast, Pluto's Neptune-orbiting cousin, Triton, exhibits only a modest visible lightcurve, a clear spectral signature of water ice, and no extensive dark regions. Yet Triton is hardly a dull or static satellite. Triton's lightcurve has recently increased in amplitude (B. Buratti, NASA Jet Propulsion Laboratory, California), its spectrum has reddened distinctly more than once, and even its atmospheric pressure has been increasing (J. Elliot, Massachusetts Institute of Technology)⁸, so there is great interest in monitoring Pluto for similar effects. As a consequence of Pluto's eccentric orbit about the Sun and the large tilt of its rotation axis, Pluto's surface is variably illuminated, and

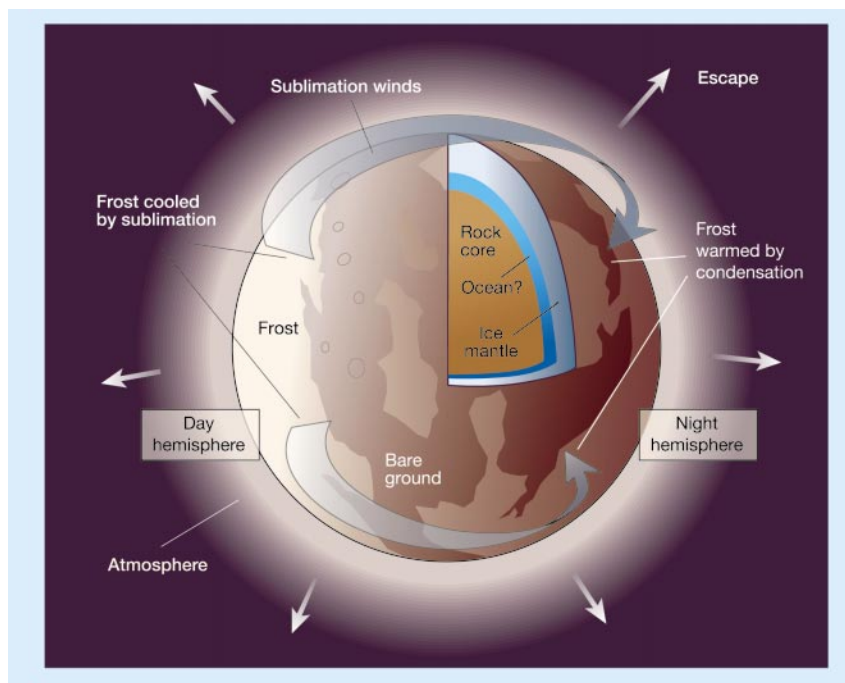


Figure 1 Defrosting Pluto. This diagram shows how volatile substances migrate across the varying terrain of Pluto's surface, and a possible internal structure for the 2,400-km-diameter planet. Frost on the sunlit hemisphere sublimates to form a vaporous atmosphere, some of which escapes but most of which flows to the dark hemisphere and condenses. The hypothesized internal structure is based on what is known about Pluto's composition and a plausible thermal history. An internal ocean is possible, though not guaranteed, as is a layer of heavy organic compounds (not shown) beneath the ocean. (Modified from a figure by J. Spencer, Lowell Observatory.)

*From Here to Pluto–Charon: A New Horizons PKB Mission Workshop, Southwest Research Institute, Boulder, Colorado, USA, 20–21 May 2002.

this, coupled with the strong lightcurve, makes it difficult to identify seasonal volatile transport or active geological effects. New lightcurve data cannot be explained by published, static models of Pluto's surface reflectance, but whether the surface of Pluto is truly changing remains to be determined (M. Buie, Lowell Observatory).

It is also fair to say that physical models of surface-atmosphere interaction have so far not done a good job of predicting what is already known of the distribution of volatiles on Pluto's surface (J. Spencer, Lowell Observatory). This complicates present efforts to work out whether Pluto's atmosphere might condense onto its surface before 2015. If it does so, as the planet retreats from the Sun, we will have missed a unique opportunity to understand the transport of volatiles on bodies like Pluto (Fig. 1), and indeed to study active atmospheric loss (D. Strobel, Johns Hopkins Univ.). In an extended Pluto atmosphere, loss processes and solar-wind interactions similar to those believed to have been important on the early Earth might occur (R. Gladstone, Southwest Research Institute, Texas).

Ongoing geological activity on Pluto is a distinct possibility. Reference was made to analogous processes across the Solar System, from tidal tectonics on Mercury and Europa (G. Collins, Wheaton College, Massachusetts) to martian surface activity, and, naturally, to the nitrogen geysers, cantaloupe terrain and cryovolcanic structures of Triton (J. Moore, NASA Ames Research Center, California). Given that volcanism and other activity on Triton is relatively geologically recent⁹, Pluto, similar in size, density and composition, may behave similarly. We have also learned from the Galileo spacecraft's flight to Jupiter that large icy satellites commonly have internal oceans¹⁰, and Pluto could well have an ocean beneath a 200-km-thick ice crust, if the conditions are right (W.B.M.).

Even Charon — half the size of Pluto — may be more geologically active than had been thought. A distinct feature in Charon's spectrum at a wavelength of 2.2 μm has been ascribed to ammonia-bearing ices¹¹. Although the reality of this feature was debated (it is not seen at all longitudes on Charon; D. Cruikshank, NASA Ames Research Center), the very presence of ammonia ice has long been equated with active geology. With any luck, however, Charon will not prove to have been so active as to have erased its ancient cratering record. Between it and Pluto, crater counts should determine the flux of impactors in the Kuiper Belt over the history of the Solar System, and thus constrain the character and flux of short-period comets from the belt and into the rest of the Solar System.

The future of Pluto-Charon science looks bright (as bright as it can be, so far away). Pluto is moving into alignment with

the Milky Way, and its first observable eclipse of a bright star since 1988 is set for this month; thereafter such eclipses occur roughly twice a year. These will provide first-class opportunities to probe Pluto's atmospheric structure. Next January will see the launch of NASA's Space Infrared Telescope Facility, which will pin down the heat budgets and thermal characteristics of terrains on Pluto and Charon, as well as determining the sizes and albedos of many other Kuiper Belt objects for comparison. And even now, the Advanced Camera for Surveys (recently installed on the Hubble Space Telescope) has begun taking the best-resolved images of Pluto yet. Nothing, however, beats the resolution of a spacecraft encounter. The fate of the New Horizons mission presently rests with the US Congress, but if it goes ahead it will provide the best answers by far to our

fundamental questions about Pluto-Charon and the Kuiper Belt. ■

William B. McKinnon is in the Department of Earth and Planetary Sciences and McDonnell Center for the Space Sciences, Washington University, St Louis, Missouri 63130, USA.

e-mail: mckinnon@levee.wustl.edu

1. Committee on Planetary and Lunar Exploration *Exploring the Trans-Neptunian Solar System* (National Academy Press, Washington DC, 1998).
2. Stern, S. A. *Sci. Am.* **286**, 56–63 (2002).
3. Stern, S. A. & Tholen, D. J. (eds) *Pluto and Charon* (Univ. Arizona Press, Tucson, 1997).
4. Douté, S. *et al. Icarus* **142**, 421–444 (1999).
5. Grundy, W. M. & Buie, M. W. *Icarus* **157**, 128–138 (2002).
6. Lellouch, E. *et al. Icarus* **147**, 220–250 (2000).
7. Young, E. F., Binzel, R. P. & Crane, K. *Astron. J.* **121**, 552–561 (2001).
8. Elliot, J. L. *et al. Icarus* **148**, 347–369 (2000).
9. Stern, S. A. & McKinnon, W. B. *Astron. J.* **119**, 945–952 (2000).
10. Zimmer, C., Khurana, K. K. & Kivelson, M. G. *Icarus* **147**, 329–347 (2000).
11. Brown, M. J. *Annu. Rev. Earth Planet. Sci.* **30**, 307–345 (2002).

Reproductive biology

Do the locomotion

After animals have mated, sperm in the female reproductive tract must race to gain that incomparable prize, fertilization of an egg. Because sperm express little of their genetic make-up in their outward appearance, it is difficult to select the good from the bad. So it might be better for an individual's sperm to cooperate rather than compete with one another. For species in which females mate with multiple partners, this will be particularly true if the sperm of one male could unite to defeat those of its rivals.

Elsewhere in this issue, Harry Moore and colleagues describe an amazing example of such altruistic behaviour in the sperm of the common European wood mouse, *Apodemus sylvaticus* (*Nature* **418**, 174–177; 2002). They find that hundreds or thousands of sperm link hooked structures on their heads and swim en masse in a train, which enables them to progress at almost twice the speed of a single sperm. These trains must break up before fertilization, so many of the component sperm commit genetic hara-kiri by undergoing a premature 'acrosome reaction'. This involves the release of enzymes that break down cell adhesion molecules,



which also makes it impossible for the sperm concerned to fertilize the egg. Somewhere on the train — perhaps it's the locomotive driver up front — there must be one acrosome-intact sperm that has retained its capacity to perform fertilization.

It might be no accident that the wood mouse is a supreme sexual performer among rodents, with males scrambling to mate polygamously with any and every promiscuous female, and they have relatively large testes to prove the point — as the picture shown here attests. So an individual male not only tries to ensure his reproductive success through sperm collaboration, as Moore *et al.* show, but also by sheer sperm numbers.

The story does not end there. Sperm motility is ultimately driven by the engine of mitochondrial DNA in the sperm's midpiece. An exciting

paper by Matthew Anderson and Alan Dixon (*Nature* **416**, 496; 2002) has shown that, in primates, the volume of the sperm midpiece is highly correlated with relative testicular size and mating behaviour, the most sexually athletic species having the largest mitochondrial midpieces to power their sperm. So we can look forward to further work to see whether the wood mice have vast mitochondrial midpieces to power their sperm trains, and whether the sperm of especially promiscuous primates such as chimpanzees would leave their human counterparts for dead in the Olympic swimming pool.

Roger V. Short

Roger V. Short is in the Department of Obstetrics and Gynaecology, University of Melbourne, Victoria 3053, Australia.
e-mail: r.short@unimelb.edu.au

Biodiversity

Bacterial game dynamics

Martin A. Nowak and Karl Sigmund

Studies of three bacterial strains engaged in an interaction that mimics the game 'rock–paper–scissors' show the importance of localized interactions in maintaining biodiversity.

It is not surprising that games as absorbing as bridge and chess have their world federations and international unions. But not everyone knows that even a game as lowly as rock–paper–scissors has its own society. This game, which must surely be very old, can be explained to any toddler. Two players signal, on a given cue, either rock (fist), paper (flat hand) or scissors (two fingers). If I display a flat hand and you show me your fist, I win, as 'paper wraps rock'. Similarly, scissors cuts paper, and rock smashes scissors. If both players make the same signal, the game ends in a draw. And in case you think of it as a rather simple-minded pastime, you should take a look at the home page of the World RPS Society¹, which is a treat. Among other features there are links to learned papers, although you are advised not to visit the links in the probabilistic section, filled as it is with "pseudo-scholastics". No such ban appears against the link to a paper in *Nature* describing three mating strategies of the male lizard *Uta stansburiana*². And now *Nature* should

hit the website again, with the report by Kerr and colleagues on page 171 of this issue³.

Kerr *et al.*³ set out to investigate the mechanisms that maintain biodiversity in ecosystems, by studying several diverse strains of *Escherichia coli* bacteria. These strains can produce a toxin, or not; and they can be resistant to the toxin, or not. We may assume that the bacterial devices for producing both the toxin and the 'antidote' that confers resistance are costly in the sense that they require resources that could otherwise have been used by the bacteria to multiply faster.

There are four potential strains. The one producing the toxin but not the antidote effectively commits suicide. This strain is a non-starter, and we may ignore it (as did Kerr *et al.*). The other three are engaged in a rock–paper–scissors type of competition. The poison- and antidote-producing strain kills that which produces neither poison nor antidote. The strain that produces the antidote but not the poison outgrows the one

that produces both, by economizing on the cost of an ineffective poison. And in the absence of the toxin-producing strain, the strain that produces no antidote outgrows the antidote-producing type, which is paying for an unneeded device.

These two-way bacterial interactions have been described previously. And similar rock–paper–scissors cycles of spiteful measures and costly countermeasures occur in other evolutionary contexts, for instance in the genetics of sexual species. Some chromosomes acquire mutations that prevent their opposite number (inherited from the other parent) from making their way into eggs or sperm, and so to the next generation. Some of these mechanisms for subverting the fair segregation of chromosomes act like the bacteria, by means of a poison-and-antidote-type principle⁴.

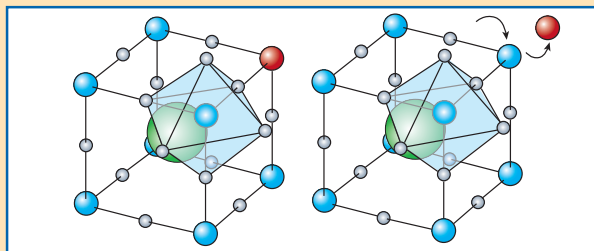
But what happens when the three *E. coli* strains are all in the same environment? The mere knowledge that the outcomes of pair-wise competition form a rock–paper–scissors cycle is not enough to predict what happens when all three types are present^{5,6}. The three competitors might co-exist permanently; this seems to apply, for instance, to the male lizards that have three different mating strategies². Or one type might be ousted, and a second type outcompeted by the third, leading to just one survivor. Kerr *et al.* find that this latter outcome holds for our bacteria. If all three strains are equally

Chemistry

Cleaning up catalysts

Since its introduction more than 20 years ago, the catalytic converter has sharply reduced automotive emissions. Using catalysts containing palladium, platinum and rhodium, a converter breaks down harmful carbon monoxide, nitrogen oxides and hydrocarbons before the exhaust leaves the car's tailpipe. But these precious metals are expensive and are produced through the intensive and polluting chemical processing of sulphide ores extracted from often treacherous underground mines.

On page 164 of this issue, Yasuo Nishihata *et al.* (*Nature* **418**, 164–167; 2002) propose a new catalyst for automotive-emissions control that lasts longer and accomplishes its task more effectively than conventional catalysts. The material, a perovskite containing small amounts of palladium (Pd), could reduce by 70–90% the amount of precious metals needed to meet today's car emission standards.



Catalytic converters consist of a highly porous ceramic structure coated with finely divided catalytic material. To ensure immediate action when the car is started, converters are placed close to the hottest part of the car, the engine. Over time, heat exposure causes the tiny particles of precious metal to agglomerate, thus reducing the catalyst's overall surface area and hence its activity. To counter this effect, conventional catalytic converters are loaded with an excess of precious metal, ensuring that performance targets are met for vehicle use over the expected range, usually 80,000 km.

Nishihata *et al.* demonstrate that excess-metal loading is not needed if the perovskite $\text{LaFe}_{0.57}\text{Co}_{0.38}\text{Pd}_{0.05}\text{O}_3$ is used as the catalyst. This material, first investigated for catalytic-converter applications in the 1970s, maintained its high metal dispersion and high catalytic activity during a 100-hour test in engine exhaust. In the same test, the activity of a conventional catalyst (alumina impregnated with palladium) decreased by 10%.

The resilience against metal-particle agglomeration results from the perovskite's ability to respond structurally to the fluctuations in

exhaust-gas composition that occur in modern petrol engines. These fluctuations switch the exhaust environment continuously from an oxidizing to a reducing atmosphere. In separate tests, the authors established that Pd is firmly incorporated into the perovskite lattice of the oxidized $\text{LaFe}_{0.57}\text{Co}_{0.38}\text{Pd}_{0.05}\text{O}_3$ catalyst (see figure, left). But the Pd atom (red) moves out of the structure, being replaced by an iron atom (blue), in the reduced material (see figure, right). It is this fully reversible hopping of Pd into and out of the perovskite structure that seems to suppress the agglomeration of the metal and thus the slow deactivation of the catalyst.

Although catalytic converters have already made a significant contribution to emissions control, Nishihata *et al.* show that more can still be done to address the environmental impact of car use.

Magdalena Helmer

frequent at first, the type producing antidote but no poison eliminates the others.

Kerr *et al.* also show, however, that this loss of diversity occurs only if the bacterial populations are well mixed. Otherwise, all three strains survive. This conclusion has been predicted on paper: a rush of theoretical investigations during the past decade has shown, in great generality, that if dispersal of a population is limited and its interactions with other populations are localized, then diversity is protected to a large degree^{7,8}. This holds for an astonishing variety of scenarios, for instance in epidemiological models, community ecology, plant genetics, animal behaviour, molecular evolution⁹ and game theory¹⁰. Such spatial models are usually much harder to analyse than their homogenized 'mean field' counterparts. But computer simulations warn us that, in many cases, 'mean field' can lead to wrong conclusions.

And Kerr and colleagues are not the first to show that localized interactions of the rock-paper-scissors type can turn a 'one winner' outcome into a dynamic coexistence of all three types, endlessly chasing each other across the board¹¹. The beauty of their paper is that they show this not only on a computer screen, but also in 'real life'. To set up the well-mixed case, the authors put all three strains in a flask, shake this cocktail, transfer a few drops to another flask, shake it again, and so on. Soon the flasks contain only the resistant, non-toxic strain. To ensure localized interactions, on the other hand, Kerr *et al.* spread the strains on a plate to let them grow, then press a cloth on the plate, transfer whatever clings to the cloth onto another plate, and so on. All three strains survive, with the boundaries between them shifting to and fro, reflecting the cyclic invasion and displacement of one strain by the next. So the outcomes on the plate and in the flask are strikingly different.

This approach opens new vistas for understanding how biological communities are built up — one of the most intriguing aspects of the study of biodiversity. The results of sequential invasions and extinctions of species can create complex links and webs in an ecosystem, not least because the outcome of an invasion depends so much on its timing and other contingencies. Large-scale experiments in community construction are generally hard to come by — not often do volcanic rocks emerge, providing barren ground for colonization. So ecologists have increasingly turned, since G. F. Gause's work in the 1930s, to manipulating mini-worlds inhabited by microbial species¹². The paper by Kerr *et al.* gives a new impetus to such investigations, by stressing the importance of the geometry of neighbourhoods. Many habitats resemble the surface of a pizza more than a well-stirred bowl of soup. ■

Martin A. Nowak is at the Institute for Advanced Study, Princeton, New Jersey 08540, USA.

e-mail: nowak@ias.edu

Karl Sigmund is at the Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Vienna, and the Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria.

e-mail: karl.sigmund@univie.ac.at

1. <http://www.worldrps.com/archive/index.html>
2. Sinervo, B. & Lively, C. M. *Nature* **380**, 240–243 (1996).
3. Kerr, B., Riley, M. A., Feldman, M. W. & Bohannan, B. J. M. *Nature* **418**, 171–174 (2002).
4. Hartl, D. L. *The Principles of Population Genetics* (Sinauer,

Sunderland, Massachusetts, 1980).

5. Hofbauer, J. & Sigmund, K. *Evolutionary Games and Population Dynamics* (Cambridge Univ. Press, 1998).
6. Weissing, F. J. in *Game Equilibrium Models I* (ed. Selten, R.) 29–96 (Springer, Berlin, 1991).
7. Durrett, R. & Levine, S. *Theor. Pop. Biol.* **46**, 363–394 (1994).
8. Dieckmann, U., Law, R. & Metz, J. A. J. (eds) *The Geometry of Ecological Interactions: Simplifying Spatial Complexity* (Cambridge Univ. Press, 2000).
9. Boerlijst, M. C. & Hogeweg, P. *Physica D* **48**, 17–28 (1991).
10. Nowak, M. A. & May, R. *Nature* **359**, 826–829 (1993).
11. Nowak, M. A., Bonhoeffer, S. & May, R. M. *Int. J. Bifurc. Chaos* **4**, 33–56 (1994).
12. Weatherby, A. J., Warren, P. H. & Law, R. *J. Anim. Ecol.* **67**, 554–566 (1998).

Apoptosis

Repulsive encounters

Giovanna Chimini

In multicellular organisms, cells are often required to die. They are then eaten by 'phagocytic' cells. But how do the phagocytes distinguish between dead and living prey? New work provides an unexpected answer.

Making decisions on the basis of social cues is an everyday challenge for human beings — and for our cells. On page 200 of this issue, for instance, Brown and colleagues¹ describe how predatory immune cells decide whether or not to eat other cells on the basis of a molecular 'handshake'.

This decision is made in the context of the programmed deletion of cells, or 'apoptosis', that is an essential feature of embryonic development, tissue organization and cell turnover in multicellular organisms^{2,3}. The programme basically consists of two tightly coupled events, the death of cells and their burial — or rather, their consumption by the

body's professional predators, immune cells known as phagocytes. Three keywords govern the burial rite⁴: swiftness, because no cellular corpses are allowed to linger around; efficiency, because the day-to-day burden of dead cells is enormous; and discretion, to protect the surrounding environment from harm. But between death and burial lies decision-making: how do phagocytes discriminate between dead or dying cells and their healthy neighbours?

It is generally accepted that the onset of the death programme makes cells appetizing because it exposes attractive 'eat me' signals⁵. These involve changes in the cell surface, which can be tracked experimentally by the

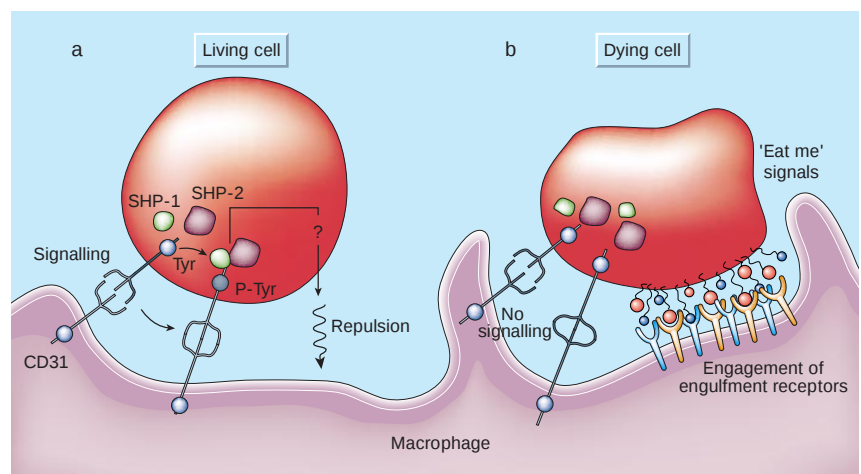


Figure 1 To eat or not to eat? In vertebrates, predatory macrophage cells continuously meet potential 'prey' and make contact with them through their respective CD31 proteins. **a**, When the target cells are healthy and viable, 'outside-in' signalling is triggered, probably involving the modification of tyrosine amino acids in the intracellular tail of CD31 with phosphate (P-Tyr), and interaction with the proteins SHP-1 and SHP-2. This results in 'inside-out' signalling, repelling the macrophage. **b**, When the target cells are dead or dying, outside-in signalling through CD31 is disabled and the interaction with the macrophage persists. This assists definitive recognition of the dying prey by specialized engulfment receptors on the macrophage.

relocation of a particular lipid — phosphatidylserine — from the inner to the outer leaflet of the plasma membrane⁶. It is also envisaged that death induces the inappropriate surface localization of molecules from internal cellular compartments⁷, which seems plausible given the violent membrane remodelling and changes in cell volume or density that occur as cells die. But none of the latter, potentially attractive, molecular signals has been identified.

Brown *et al.*¹ now provide evidence of an alternative sorting system that is based on negative, rather than positive, cues. They propose that repulsive impulses from healthy cells actively discourage the predatory phagocytes. But these impulses are disabled during cell death, and it is this that allows prolonged contact between predator and prey. Such long-lasting interactions assist the definitive recognition of attractive signals generated later on the dying cells by specialized receptors on the phagocyte surface⁸.

The authors obtained clues to this mechanism by reconstituting, in a flow-chamber system, encounters between macrophages (a particularly efficient type of phagocyte) and leukocytes (a potential prey). Here, and consistent with previous observations made by video recording⁹, macrophages methodically palpate any leukocytes that they encounter. If a leukocyte is healthy, it quickly detaches from the macrophage. But if it is dead or dying, the contact lasts long enough to allow the cell to be eaten.

What is the molecular basis for this interaction? Adhesion receptors are proteins that allow cells to contact other cells or their extracellular matrix, and are among the best candidates for the molecules that mediate the link between macrophages and dying cells, as well as the switch between attraction and repulsion. Brown *et al.* identified CD31 as the adhesion receptor in this case. CD31 is a widely expressed membrane protein that can be involved in a variety of interactions¹⁰. It binds to several different partners to modulate adhesiveness and cell activation, and can also interact with CD31 molecules on the surface of other cells. For instance, CD31 on leukocytes can bind to CD31 on cells of endothelial tissue layers, allowing the leukocytes to migrate across the endothelia towards a site of infection¹¹. Here, a repeated shift from attachment to detachment — in response to bidirectional signalling through CD31 — is a key determinant of cell progression.

So what happens in our cellular life-or-death situation? Brown *et al.* first show that the predominant method by which leukocytes are tethered to macrophages involves a 'handshake' between their respective CD31 proteins. They also forced cells that do not normally express CD31 to do so and found that, when these cells were healthy, they first

attached to and then detached from macrophages. But when they were dying, they remained attached. Brown *et al.* also engineered living leukocytes to express a mutant CD31 that could not anchor to the membrane or could not transmit signals into the leukocyte on binding to CD31 on macrophages. These leukocytes also remained attached. Other data hinted that CD31 signalling is naturally disabled in dying leukocytes, preventing detachment.

Thus, the idea is that when a healthy cell encounters a macrophage, signalling through the bound CD31 receptors leads to the cell's being repelled and detaching (although the molecular details of repulsion are not known). But in dying cells, signalling through CD31 is somehow disrupted; the cell is not repelled, and the balance is shifted towards its engulfment (Fig. 1).

Although it is not yet known how the CD31-mediated signal is disabled during cell death, this view of the burial procedure is immediately appealing. It seems intuitive and economically logical to associate death with the inactivation of pre-existing signalling circuits, rather than with the emergence of new molecular pathways. Moreover, the inactivation of existing circuits might be quicker than building up new pathways, which fits with the finding that patrolling macrophages select their prey well before any massive cellular modifications occur.

If the inactivation of repulsive signalling pathways is indeed an early event in cell death, then it might shed new light on studies of the nematode worm *Caenorhabditis elegans* in which cells that were 'almost alive' were eliminated^{12,13}. These events have been interpreted as evidence that phagocytes can actively kill their prey, rather than simply consuming dead cells. But an alternative explanation is that a loss of repulsive sig-

nalling, and hence an attachment to phagocytes, occurs before the key apoptotic events — that is, before final commitment to cell death. In other words, phagocytes cross a 'point of no return' in their decision to eliminate dying cells before those cells are irrevocably committed to the decision to die. That suggestion is bolstered by the finding that CD31 does not seem to be destroyed by the apoptotic protein-degrading enzymes, hinting that the switch from repulsive to attractive signalling might become independent of definitive activation of the death machinery.

It is still possible, however, that negative sorting does not occur in invertebrates such as *C. elegans*, instead representing an upgrading of the death programme to vertebrates. Such fine-tuning might have emerged to deal more efficiently with the varied situations of cell death seen in more complicated multicellular organisms.

Giovanna Chimini is at the Centre d'Immunologie de Marseille-Luminy, INSERM-CNRS-Université de la Méditerranée, Parc Scientifique de Luminy, Case 906, 13288 Marseille, France.
e-mail: chimini@ciml.univ-mrs.fr

1. Brown, S. *et al.* *Nature* **418**, 200–203 (2002).
2. Raff, M. *Nature* **396**, 119–122 (1998).
3. Kerr, J. F. R., Wyllie, A. H. & Currie, A. R. *Br. J. Cancer* **26**, 239–257 (1972).
4. Savill, J. & Fadok, V. *Nature* **407**, 784–788 (2000).
5. Savill, J., Fadok, V., Henson, P. & Haslett, C. *Immunol. Today* **14**, 131–136 (1993).
6. Fadok, V., Bratton, D. L., Frasch, S. C., Warner, M. & Henson, P. M. *Cell Death Differ.* **5**, 551–562 (1998).
7. Hart, S. P., Ross, J. A., Ross, K., Haslett, C. & Dransfield, I. *Cell Death Differ.* **7**, 493–503 (2000).
8. Fadok, V. A., Bratton, D. L. & Henson, P. M. *J. Clin. Invest.* **108**, 957–962 (2001).
9. Parnai, R., Raff, M. C. & Scholes, J. *Curr. Biol.* **10**, 857–860 (2000).
10. Newman, P. J. *J. Clin. Invest.* **99**, 3–8 (1997).
11. Muller, W. A., Weigl, S. A., Deng, X. & Phillips, D. M. *J. Exp. Med.* **178**, 449–460 (1993).
12. Reddien, P. W., Cameron, S. & Horvitz, H. R. *Nature* **412**, 198–202 (2001).
13. Hoepfner, D. J., Hengartner, M. O. & Schnabel, R. *Nature* **412**, 202–206 (2001).

Theoretical biology

Ants on a Turing trail

Peter Hammerstein and Olof Leimar

The behaviour of ants when dealing with their dead has parallels with biological pattern formation more generally, for instance as seen during development.

Social insects can build complex structures such as termite mounds that resemble skyscrapers, complete with air conditioning. But unlike human builders, they work without the help of an architect or a blueprint. This is similar to the process of morphogenesis, by which cells of a developing organism produce features on a scale much larger than their own size. The cells do not possess an internal representation of these features, be it the pattern on a seashell or the basic segmentation of a fruitfly.

Local activation and long-range inhibition^{1–3} have traditionally been regarded as important characteristics of pattern formation, the general idea being that an elevated local concentration of a pattern-forming substance should induce further local build-up, and that a high local concentration in one place should also inhibit a build-up some distance away. There are few cases in biology where the actual mechanisms have been demonstrated. But, writing in *Proceedings of the National Academy of Sciences*, Theraulaz

*et al.*⁴ present an example of such a process in social insects. Their careful experimentation has revealed how the behaviour of individual workers of the ant *Messor sancta* produces spatial patterns in a colony's disposal of corpses — the so-called ant cemeteries.

In the first series of Theraulaz and colleagues' experiments, the ants had access to a circular arena in which ant corpses were homogeneously distributed along the perimeter. The ants had a strong tendency to follow the wall of the arena — so this is effectively a one-dimensional situation, which facilitates mathematical analysis of the process. The authors monitored the formation of clusters of dead ants and the dynamics of their spatial distribution. The result that emerged was puzzling. Instead of quickly choosing one or a few fixed locations for piles of corpses, the ants formed many clusters, some of which grew while others disappeared after some effort had already been made to build them. The number of clusters first grew, reaching a maximum after three hours. Later it decreased and remained constant when a stable spatial pattern was finally established. Subsequent experiments were performed to estimate the parameters of individual ant behaviour, such as the probabilities that a given ant would pick up or drop a corpse, or would make a U-turn while carrying a corpse along the perimeter. For this purpose the authors manipulated cluster size.

The observations indicate that a mechanism of short-range activation and long-range inhibition is at work. The activation would consist of a behavioural tendency to drop corpses with a probability that increases with the density of corpses in the immediate neighbourhood. This mechanism leads to local amplification of corpse density, unless it is overridden by the inhibitory process. Inhibition occurs through the ants' tendency to pick up corpses and carry them for considerable distances, leading to the long-range depletion of the clusters.

Theraulaz *et al.* developed a mathematical model for the process of ant cemetery construction. This model is similar to one by Gierer and Meinhardt² that has been applied to seashell pattern formation³. The ant cemetery model explained not only the final formation of a stable distribution of large piles but also the intermediate occurrence of clusters that then disappear. For the dynamics of how the average number of clusters changes with time and when its maximum is reached, the model predictions were in close agreement with the experiments. Furthermore, as the inhibition–activation structure of the model suggests, cluster formation did not occur if the initial density of corpses was too low. Results such as these show that pattern formation can be understood at a deep level and that the behavioural mechanisms behind it are very simple.

What general insights do we gain about the ants? Given that their behaviour followed the mathematical model so closely, it seems as though they have no mental concept of the large-scale process in which they participate. When they move corpses in the cemetery, their actions are guided mainly by aspects of the local temporal and spatial environment, and possibly by chance, but not by foresight and planning. One could then say that the structures simply appear as a by-product of the ants' shortsighted way of dealing with corpses. Would this be adaptive in the evolutionary sense? At first glance the ants seem to act inefficiently. But it might well be a better use of a small brain not to attempt to be an excellent architect. The robustness of the procedure that structures the cemetery could outweigh a certain waste of time and energy.

Ants in the wild do not have access to a circular arena, but the same mechanisms that lead to periodic patterns in the arena are likely to be operating. The structures of the natural cemeteries of *M. sancta* have not been studied systematically, so it is not known whether they sometimes consist of several piles or whether the ants' behavioural mechanisms usually produce a single pile. Nevertheless, the experiments with the circular arena reveal how simple mechanisms can act to create structure in a controlled and simple world for which it is easy to compare theoretical and empirical findings.

This is notable because the results reported by Theraulaz *et al.*⁴ are likely to have parallels in other collective processes. For instance, decisions in morphogenesis are made by cells, which are difficult to envisage in the role of architects. Organisms and 'superorganisms' of social insects thus share a problem, so ants might teach us valuable lessons about morphogenesis. The potential of this analogy was recognized 25 years ago by Deneubourg⁵. We now know that various cases of patterns on seashells³ and the formation of piles in ant cemeteries might be explained by similar underlying principles. These principles exemplify the basic conceptual model of activation and inhibition that Gierer and Meinhardt² created to make Turing's¹ abstract ideas about morphogenesis amenable to observation and experiment. The ants may indeed be on a Turing trail. ■

Peter Hammerstein is at the Institute for Theoretical Biology, Humboldt University, Invalidenstrasse 43, D-10115 Berlin, Germany.

e-mail: p.hammerstein@biologie.hu-berlin.de

Olof Leimar is in the Department of Zoology, Stockholm University, SE-106 91 Stockholm, Sweden.

1. Turing, A. *Phil. Trans. R. Soc. Lond. B* **237**, 37–72 (1952).
2. Gierer, A. & Meinhardt, H. *Kybernetik* **12**, 30–39 (1972).
3. Meinhardt, H. *The Algorithmic Beauty of Sea Shells* (Springer, Berlin, 1995).
4. Theraulaz, G. *et al. Proc. Natl Acad. Sci. USA* Early edition 10.1073/pnas.152302199 (8–12 July, 2002).
5. Deneubourg, J. L. *Insectes Sociaux* **2**, 117–130 (1977).



100 YEARS AGO

A few examples of the practical application of scientific education in Germany are given in the *Journal* of the Society of Arts. The sugar industry is the first illustration of the progress of industry through science. In 1840, 154,000 tons of beetroot were crushed, from which 8000 tons of raw sugar were produced, showing about 5½ per cent. of raw sugar extracted from the root. Twenty years later, 1,500,000 tons were treated which produced 128,000 tons of sugar, or about 8 per cent. Last year about 12,000,000 tons were crushed, which produced 1,500,000 tons of raw sugar, raising the percentage to 13. This advance is due entirely to scientific treatment. The production of dry colours, chemicals and dyes in Germany shows a corresponding increase in production and dividend-paying capacity... A great advance has also been made in the scientific instrument industry. The value of the exports from Germany of scientific instruments in the year 1898 was about 250,000/ — three times what it was in 1888 — and the work gave employment to 14,000 people.

From *Nature* 10 July 1902.

50 YEARS AGO

The Mitotic Cycle. By Dr. Arthur Hughes. It is rightly emphasized on the dust-cover of this valuable monograph that the process of cell division presents one of the most difficult problems the experimental biologist has yet attempted to solve, and if this present account of the mitotic cycle is not an easily flowing and well-balanced narrative, the reflexion is not on the author but on the present state of our knowledge of the subject. There is a large and widely scattered literature of unequal relevance and of uneven quality. The lines of advance have been largely dictated by considerations of the materials and techniques available, and a strong medical bias is also evident. Thus we now have an extensive knowledge of the early cleavage of a certain few eggs, of the growth in culture media of a certain few tissues, of the methods of induction of cancerous growths, and of the methods of mitotic inhibition by a multitude of diverse substances. The obvious questions posed by the mitotic activity of normal animal and plant tissues have been almost entirely neglected, although very recently a start has been made towards their solution.

From *Nature* 12 July 1952.

Snake circumvents constraints on prey size

Instead of swallowing a victim whole, crab-eating snakes tear off bite-sized pieces.

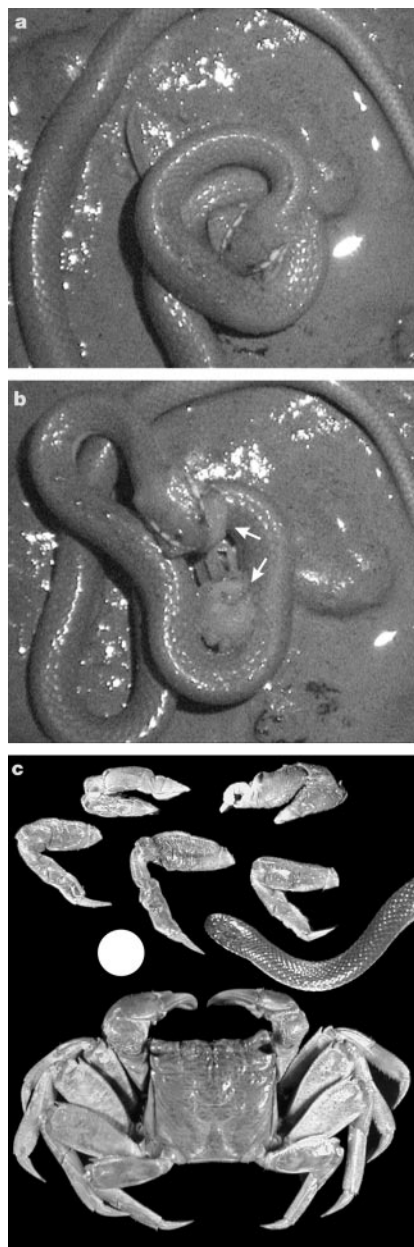


Figure 1 Feeding behaviour and prey size of the snake *Gerarda prevostiana*. **a**, **b**, Infrared video images of *G. prevostiana* 12.6 s before (**a**) and 0.6 s after (**b**) completely tearing apart the carapace of a freshly moulted grapsid crab that was considerably wider than the snake's head. Both pieces of the carapace (arrows) were consumed. For videos showing this behaviour, see www.biology.uc.edu/faculty/jayne/gerarda_feed2.htm. **c**, Five appendages (top) from a freshly moulted crab (*Episesarma versicolor*) that was attacked and partially consumed by a 33-g *G. prevostiana* snake (middle) in the wild. The intact, 41-g *E. versicolor* shown at the bottom is similar in size to the crab that was attacked. The white circle (actual diameter, 13 mm) is the maximum gape of the snake (determined by inserting cylindrical gauges with 1-mm diameter increments into the mouth of the anaesthetized snake).

For animals who are unable to take bites out of their food, the size of the food item that can be consumed is constrained by the maximal size of the mouth opening (gape) — snakes are an example of gape-limited predators and they usually swallow their prey whole^{1,2}. Here we describe unique feeding behaviours in two closely related species of snake, which circumvent their gape limitation by removing and consuming pieces from newly moulted crabs that are too large to be swallowed intact. This evolutionary innovation is surprising, as the needle-like teeth and highly mobile bones that facilitate the capture and engulfment of large, whole prey by snakes are ill-suited both to cutting and to generating large bite forces.

We captured individuals of the homalopsine snake species *Gerarda prevostiana* and *Fordonia leucobalia* and examined their gut contents. We found that *G. prevostiana* contained the remnants of only freshly moulted crabs, whereas *F. leucobalia* contained material exclusively from hard-shelled crabs. *F. leucobalia* has several morphological³ and some behavioural⁴ specializations that are associated with its unusual diet. However, *G. prevostiana*, the diet and feeding behaviour of which have not previously been investigated, lacks the hypertrophied cranial musculature and short and blunt teeth of *F. leucobalia*⁵. We therefore collected *G. prevostiana* and a variety of crabs in Singapore and observed the snakes feeding in a dark room with infrared video cameras.

Captive *G. prevostiana* ($n = 15$) refused to attack hard-shelled crabs, but bit and rapidly swallowed freshly moulted crabs. In 85% of trials ($n = 26$), the snake pulled its head through a loop in its body and continued to pull on the crab while coiled around the animal (Fig. 1a, b). In some trials (35%), snakes broke off and consumed up to four crab legs but none of the carapace. The carapace of the crab was either partially or completely torn apart in 20% and 32%, respectively, of the 22 feedings in which snakes demonstrated the 'loop-and-pull' behaviour (Fig. 1a, b). Tearing of the carapace was most common for large crabs, and only occurred when a loop-and-pull action was used.

We also captured two *G. prevostiana* individuals that had consumed pieces of newly moulted crabs that were much larger relative to the snake than any of those used in laboratory trials (Fig. 1c). Consequently, ripping crabs apart is unlikely to be incidental, and probably allows *G. prevostiana* to regularly consume crabs with body sizes that exceed our estimate of its maximal gape.

These two species of crab-eating snake have not merely retained a common feeding mechanism of a crustacean-eating ancestor. The feeding mechanism used by captive *F. leucobalia* ($n = 12$) to eat both hard-shelled ($n = 50$) and soft-shelled ($n = 6$) crabs differed from that used by *G. prevostiana* in the mode of attacking prey, how prey was restrained, the usual orientation for swallowing the crab's body, and how pieces were torn from prey. In contrast, the patterns of coiling shown by many lineages of constricting snakes are usually evolutionarily conserved⁶. When *F. leucobalia* consumed hard-shelled crabs, the carapace always remained intact, contrary to a suggestion that this snake uses its jaws to crush crabs³. By using a method similar to that used for breaking legs off crabs⁴, however, captive *F. leucobalia* did tear carapaces apart in 50% of trials with very large, soft-shelled crabs.

We are unaware of any other species of snake that regularly tears its prey apart so that it can consume larger prey than could be swallowed whole. The novel feeding mechanism used by *G. prevostiana* is surprising in view of its relatively unremarkable anatomy, which cautions against drawing functional inferences from anatomy in the absence of behavioural studies.

The diversity of tropical aquatic snakes in Singapore correlates well with their range of feeding mechanisms and specialisms. Within a single mangrove, *F. leucobalia* and *G. prevostiana* consume the same species of crab, but partition this resource by relying on different moult stages; two other species, *Cantoria violacea* and *Cerberus rynchops*, eat only snapping shrimp and fish, respectively⁷.

Bruce C. Jayne*, **Harold K. Voris†**,
Peter K. L. Ng‡

*Department of Biological Sciences, University of Cincinnati, Ohio, 45221-0006, USA
e-mail: jaynebc@email.uc.edu

†Division of Amphibians and Reptiles, Field Museum of Natural History, 1400 South Lake Shore Drive, Chicago, Illinois 60605, USA

‡Raffles Museum of Biodiversity Research, Department of Biological Sciences, National University of Singapore, Science Drive 4, Singapore 119260

1. Cundall, D. & Greene, H. W. in *Feeding: Form, Function and Evolution in Tetrapod Vertebrates* (ed. Schwenk, K.) 293–333 (Academic, New York, 2000).
2. Arnold, S. J. in *Snakes: Ecology and Behavior* (ed. Collins, J. T.) 87–115 (McGraw-Hill, New York, 1993).
3. Savitzky, A. H. *Am. Zool.* **23**, 397–409 (1983).
4. Shine, R. & Schwaner, T. *Copeia* **1985**, 1067–1071 (1985).
5. Voris, H. K. et al. *Copeia* (in the press).
6. Greene, H. W. & Burghardt, G. M. *Science* **200**, 74–77 (1978).
7. Voris, H. K. & Murphy, J. C. *J. Nat. Hist.* (in the press).

Competing financial interests: declared none.

Recombination

Multiply infected spleen cells in HIV patients

The genome of the human immunodeficiency virus is highly prone to recombination^{1–3}, although it is not obvious whether recombinants arise infrequently or whether they are constantly being spawned but escape identification because of the massive and rapid turnover of virus particles^{4,5}. Here we use fluorescence *in situ* hybridization to estimate the number of proviruses harboured by individual splenocytes from two HIV patients, and determine the extent of recombination by sequencing amplified DNA from these cells. We find an average of three or four proviruses per cell and evidence for huge numbers of recombinants and extensive genetic variation. Although this creates problems for phylogenetic analyses, which ignore recombination effects, the intracellular variation may help to broaden immune recognition.

The recombination rate of HIV-1 is roughly three events per genome for every round of replication, with a range of one to seven crossovers¹ (which is about tenfold greater than its point-substitution rate of about 0.25 per genome per round⁶). Some strains in global circulation are composites of viruses from two or three different clades^{7,8}. If the majority of infected cells harbour a single provirus, then recombinants will be detectable, albeit only rarely. However, if most cells contain multiple proviruses, particularly if they are genetically distinct, then budding viruses will encapsulate genetically distinct genomes that will recombine in the next round of infection, thereby generating complex mosaic structures.

We used fluorescence *in situ* hybridization (FISH) to quantify the number of proviruses per cell, and tested their sequence divergence after amplification by polymerase chain reaction (PCR) of DNA from single splenocytes taken from two HIV-1-infected patients, B and R (Fig. 1).

Frozen cells were thawed and positively selected by magnetic beads coated with anti-CD4 antibody, which yielded about 94% CD4⁺ T cells by flow-assisted cell sorting (FACS) analysis. For both B and R, more than 100 HIV-1-positive interphase cells were studied (Fig. 1a). The proviral copy number ranged from one to eight per cell, with a mean of around 3.2 (Fig. 1b). The frequency distributions were remarkably similar, despite different clinical presentations and viral loads.

With 75–80% of HIV-1 infected cells *in vivo* harbouring two or more proviruses, recombinants should be readily identifiable. We laser-dissected single, HIV-positive, interphase nuclei and transferred them to

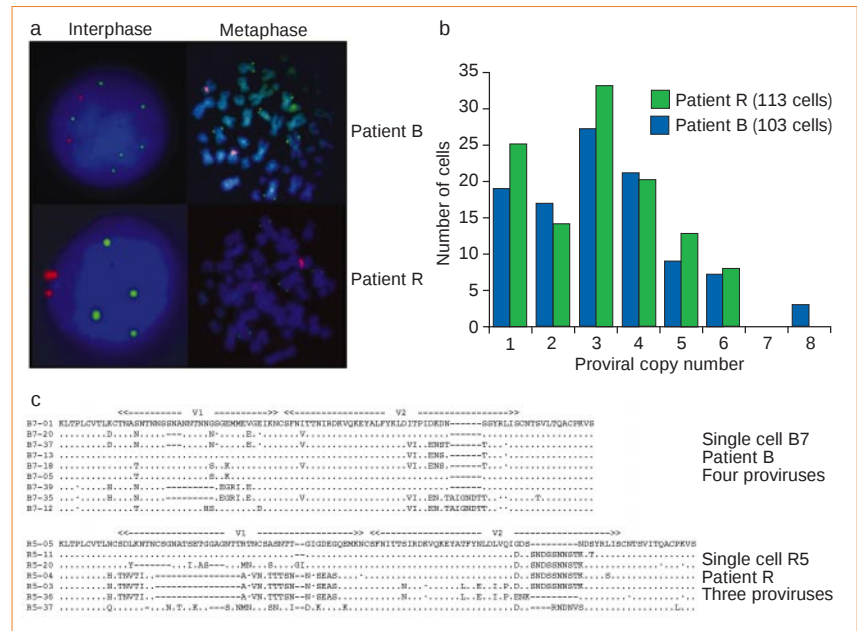


Figure 1 HIV-1 infected splenocytes harbour numerous diverse proviruses. **a**, HIV-specific fluorescence *in situ* hybridization of HIV-1 proviruses in interphase nuclei and metaphase chromosomes from splenocytes. Green spots correspond to integrated HIV genomes, or proviruses; red represents the centromeric region of chromosome 12. Patient B, clinical stage B1: blood CD4, 583 cells per μ l; plasma viral load, 5,900 RNA copies per ml; patient R, stage C2: blood CD4, 317 cells per μ l; plasma viral load, 126,000 RNA copies per ml. **b**, Frequency distribution of the number of HIV-1 proviral copies per infected cell. **c**, V1V2 Env sequence diversity within single infected CD4⁺ T cells. Only the differences are shown; dashes indicate gaps, dots denote synonymous substitutions.

PCR-assay tubes. We chose the hypervariable V1V2 region of the HIV-1 *env* gene for amplification because it is one of the most variable regions of the HIV-1 genome and thus offers the greatest resolution. The nested PCR protocol used was sensitive to about one copy⁹. To assess the possibility that HIV DNA could leak from HIV-positive cells, we microdissected numerous HIV-negative interphase nuclei and subjected them to PCR. No positive reactions were observed, excluding the possibility of cross-contamination between cells.

A collection of sequences derived from single cells harbouring three or four proviruses is shown in Fig. 1c. The most striking features are the numbers of distinct sequences per cell and the extent of genetic variation within a single cell — up to 34% amino-acid difference for cell R5 (compare R5-36 and R5-37). Several sequences are recombinants, for example B7-35 and R5-36. With most of the infected cells harbouring two or more genetically different proviruses, the necessary conditions for rampant generation of recombinants are fulfilled, indicating that recombination is important in shaping the evolution of HIV in individual patients.

Standard phylogenetic analyses, which ignore recombination, therefore probably overestimate branch lengths¹⁰. Moreover, mutation in an epitope that is encoded by one provirus would still leave the cell vulnerable to recognition of the same epitope encoded by the other proviruses. The

number of distinct antigenic epitopes presented by a single cell may be increased, thereby broadening immune recognition. Although massive recombination may help the virus to recover from deleterious mutations and the effects of relentless bottlenecks inherent in the chronic phase of HIV infection^{4,5,11}, the price that the virus pays for multiple infection may be a broadening of immune recognition of the infected cell.

Andreas Jung*, Reinhard Maier*, Jean-Pierre Vartanian†, Gennady Bocharov‡, Volker Jung§, Ulrike Fischer§, Eckart Meese§, Simon Wain-Hobson†, Andreas Meyerhans*

Departments of *Virology and §Human Genetics, University of the Saarland, 66421 Homburg, Germany

e-mail: andreas.meyerhans@uniklinik-saarland.de

†Unité de Rétrovirologie Moléculaire, Institut Pasteur, 75724 Paris cedex 15, France

‡Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow, Russia

1. Jetz, A. E. *et al.* *J. Virol.* **74**, 1234–1240 (2000).
2. Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. *Nature* **374**, 124–126 (1995).
3. McCutchan, F. E. *AIDS* **14**, S31–S44 (2000).
4. Ho, D. D. *et al.* *Nature* **373**, 123–126 (1995).
5. Wei, X. *et al.* *Nature* **373**, 117–122 (1995).
6. Mansky, L. M. & Temin, H. M. *J. Virol.* **69**, 5087–5094 (1995).
7. Carr, J. K. *et al.* *Virology* **247**, 22–31 (1998).
8. Hoelscher, M. *et al.* *AIDS* **15**, 1461–1470 (2001).
9. Gratten, S., Cheynier, R., Dumaourier, M. J., Oksenhendler, E. & Wain-Hobson, S. *Proc. Natl. Acad. Sci. USA* **97**, 14566–14571 (2000).
10. Schierup, M. H. & Hein, J. *Genetics* **156**, 879–891 (2000).
11. Wain-Hobson, S. *Nature* **366**, 22 (1993).

Competing financial interests: declared none.

A new hominid from the Upper Miocene of Chad, Central Africa

Michel Brunet*, Franck Guy†‡, David Pilbeam†, Hassane Taisso Mackaye‡, Andossa Likius*‡, Djimdoumbaye Ahounta§, Alain Beauvilain§, Cécile Blondel*, Hervé Bocherens||, Jean-Renaud Boisserie*, Louis De Bonis*, Yves Coppens¶, Jean Dejax#, Christiane Denys#, Philippe Dourigner☆, Véra Eisenmann#, Gongdibé Fanone§, Pierre Fronty*, Denis Geraads**, Thomas Lehmann*, Fabrice Lihoreau*, Antoine Louchart††, Adoum Mahamat§, Gildas Merceron*, Guy Mouchelin*, Olga Otero*, Pablo Pelaez Campomanes‡‡, Marcia Ponce De Leon§§, Jean-Claude Rage#, Michel Sapanet|||, Mathieu Schuster☆, Jean Sudre||, Pascal Tassy#, Xavier Valentin*, Patrick Vignaud*, Laurent Viriot*, Antoine Zazzo¶¶ & Christoph Zollikofer§§

* Faculté des Sciences et CNRS UMR 6046, Université de Poitiers, 40 Avenue du Recteur Pineau, 86022 Poitiers Cedex, France

† Peabody Museum, Harvard University, 11 Divinity Avenue, Cambridge, Massachusetts 02138, USA

‡ Université de N'Djaména, BP 1117, N'Djaména, Tchad

§ Centre National d'Appui à la Recherche, BP 1228, N'Djaména, Tchad

|| Institut des Sciences de l'Evolution, CNRS UMR 5554, Université de Montpellier II, Place E. Bataillon, 34095 Montpellier Cedex 5, France

¶ Collège de France, 3 rue d'Ulm, and # Muséum National d'Histoire Naturelle et CNRS UMR 8569, rue Cuvier, 75005 Paris, France

☆ Centre de Géochimie de la Surface, CNRS UMR 7517, Université Louis Pasteur, 1 rue Blessig, 67084 Strasbourg, France

** Centre National de Recherche Scientifique UPR 2147, 44 rue de l'Amiral Mouchez, 75014 Paris, France

†† Centres des Sciences de la Terre, CNRS UMR 5125, Université Claude Bernard, 27-43 Bd du 11 novembre 1918, 69622 Villeurbanne, France

‡‡ Museo de Ciencias Naturales, C/Gutiérrez Abascal 2, 28006 Madrid, España

§§ Anthropologisches Institut/Multimedia Laboratorium, Universität Zürich-Irchel, Winterthurer Str. 190, 8057 Zürich, Switzerland

||| Centre Hospitalier Universitaire, Université de Poitiers, rue de la Milétrie, 86021 Poitiers Cedex, France

¶¶ Centre National de Recherche Scientifique UMR 162, et Institut National de la Recherche Agronomique, Université Pierre et Marie Curie, 4 place Jussieu, 75252 Paris Cedex 05, France

The search for the earliest fossil evidence of the human lineage has been concentrated in East Africa. Here we report the discovery of six hominid specimens from Chad, central Africa, 2,500 km from the East African Rift Valley. The fossils include a nearly complete cranium and fragmentary lower jaws. The associated fauna suggest the fossils are between 6 and 7 million years old. The fossils display a unique mosaic of primitive and derived characters, and constitute a new genus and species of hominid. The distance from the Rift Valley, and the great antiquity of the fossils, suggest that the earliest members of the hominid clade were more widely distributed than has been thought, and that the divergence between the human and chimpanzee lineages was earlier than indicated by most molecular studies.

From their initial description in 1925¹ until 1995, hominids from the Pliocene (5.3–1.6 million years, Myr) and late Upper Miocene (~7.5–5.3 Myr) were known only from southern and eastern Africa. This distribution led some authors to postulate an East African origin for the hominid clade (where the term 'hominid' refers to any member of that group more closely related to extant humans than to the extant chimpanzee, *Pan*)^{2,3}. The focus on East Africa has been especially strong in the past decade, with the description of several new forms from Kenya and Ethiopia, including *Kenyanthropus platyops* (3.5 Myr; ref. 4); *Australopithecus anamensis* (3.9–4.1 Myr; ref. 5); *Ardipithecus ramidus ramidus* (4.4 Myr; ref. 6); *Ardipithecus ramidus kadabba* (5.2–5.8 Myr; ref. 7) and *Orrorin tugenensis* (~6 Myr; refs 8, 9). The discoveries of *A. ramidus ramidus*, *A. r. kadabba* and *O. tugenensis* have extended the human lineage well back into the Miocene. However, the discovery of *Australopithecus bahrelghazali* in Chad, central Africa^{10,11}, demonstrated a considerably wider geographic range for early hominids than conventionally expected.

Since 2001, the Mission Paléoanthropologique Franco-Tchadienne (MPFT), a scientific collaboration between Poitiers University, Ndjamen University and Centre National d'Appui à la Recherche (CNAR) (N'Djaména), has recovered hominid specimens, including a nearly complete cranium, from a single locality (TM 266) in the Toros-Menalla fossiliferous area of the Djurab Desert of northern Chad (Table 1). The constitution of the associated fauna suggests that the fossils are older than material dated at 6 Myr from Lukeino, Kenya^{8,9}. Preliminary comparison with the fauna from the Nawata formation at Lothagam, Kenya^{12,13}, suggests that the fossils are from

the Late Miocene, between 6 and 7 Myr old. All six recovered specimens are assigned to a new taxon that is, at present, the oldest known member of the hominid clade.

Systematic palaeontology

Order Primates L., 1758

Suborder Anthropeoidea Mivart, 1864

Superfamily Hominoidea Gray, 1825

Family Hominidae Gray, 1825

Sahelanthropus gen. nov.

Etymology. The generic name refers to the Sahel, the region of Africa bordering the southern Sahara in which the fossils were found.

Generic description. Cranium (probably male) with an orthognathic face showing weak subnasal prognathism, a small ape-size braincase, a long and narrow basicranium, and characterized by the following morphology: the upper part of the face wide relative to a mediolaterally narrow and anteroposteriorly short lower face; a large canine fossa; a small and narrow U-shaped dental arch; orbits separated by a very wide interorbital pillar and crowned with a large, thick and continuous supraorbital torus; a flat frontal squama with no supratatorial sulcus but with a marked postorbital constriction; a small, posteriorly located sagittal crest and a large nuchal crest (at least, in presumed males); a flat and relatively long nuchal plane with a large external occipital crest; a large mastoid process; small occipital condyles; a short, anteriorly narrow basioccipital; the long axis of the petrous temporal bone oriented roughly 30° relative to

the sagittal plane; the biporion line touching the basion; a round external auditory porus; a broad glenoid cavity with a large postglenoid process; a robust and superoinferiorly short mandibular corpus associated with a wide extramolar sulcus; a large, anteriorly opening mental foramen centred beneath lower teeth P_4 – M_1 , below midcorpus height; relatively small incisors; distinct marginal ridges and multiple tubercles on the lingual fossa of upper I^1 ; small (presumed male) upper canines longer mesiodistally than buccolingually; upper and lower canines with extensive apical wear; no lower C – P_3 diastema; upper and lower premolars with two roots; molars with low rounded cusps and bulbous lingual faces, M^3 triangular and M_3 rounded distally; enamel thickness of cheek teeth intermediate between *Pan* and *Australopithecus*.

Differential diagnosis. *Sahelanthropus* is distinct from all living great apes in the following respects: relatively smaller canines with apical wear, the lower showing a full occlusion above the well-developed distal tubercle, probably correlated with a non-honing C – P_3 complex (P_3 still unknown).

Sahelanthropus is distinguished as a hominid from large living and known fossil hominoid genera in the following respects: from *Pongo* by a non-concave lateral facial profile, a wider interorbital pillar, superoinferiorly short subnasal height, an anteroposteriorly short face, robust supraorbital morphology, and many dental characters (described below); from *Gorilla* by smaller size, a narrower and less prognathic lower face, no supratoral sulcus, and smaller canines and lower-cusped cheek teeth; from *Pan* by an anteroposteriorly shorter face, a thicker and more continuous supraorbital torus with no supratoral sulcus, a relatively longer braincase and narrower basicranium with a flat nuchal plane and a large external occipital crest, and cheek teeth with thicker enamel; from *Samburupithecus*¹⁴ by a more anteriorly and higher-placed zygomatic process of the maxilla, smaller cheek teeth with lower cusps and without lingual cingula, and smaller upper premolars and M^3 ; from *Ouranopithecus*¹⁵ by smaller size, a superoinferiorly, anteroposteriorly and mediolaterally shorter face, relatively thicker continuous supraorbital torus, markedly smaller but mesiodistally longer canines, apical wear and large distal tubercle in lower canines, and thinner postcanine enamel; from *Sivapithecus*¹⁶ by a superoinferiorly and anteroposteriorly shorter face with non-concave lateral profile, a wider interorbital pillar, smaller canines with apical wear, and thinner cheek-teeth enamel; from *Dryopithecus*¹⁷ by a less prognathic lower face with a wider interorbital pillar, larger supraorbital torus, and thicker postcanine enamel.

Sahelanthropus is also distinct from all known hominid genera in the following respects: from *Homo* by a small endocranial capacity (preliminary estimated range 320–380 cm³) associated with a long flat nuchal plane, a longer truncated triangle-shaped basioccipital, a flat frontal squama behind a robust continuous and undivided supraorbital torus, a large central upper incisor, and non-incisiform canines; from *Paranthropus*¹⁸ by a convex facial profile that is less

mediolaterally wide with a much smaller malar region, no frontal trigone, the frontal squama with no hollow posterior to glabella, a smaller, longer and narrower braincase, the zygomatic process of the maxilla positioned more posterior relative to the tooth row, and markedly smaller cheek teeth; from *Australopithecus*^{19–21} by a less prognathic lower face (nasospinale–prosthion length shorter at least in presumed males) with a smaller malar (infraorbital) region and a larger, more continuous supraorbital torus, a relatively more elongate braincase, a relatively long, flat nuchal plane with a large external occipital crest, non-incisiform and mesiodistally long canines, and thinner cheek-teeth enamel; from *Kenyanthropus*⁴ by a narrower, more convex face, and a narrower braincase with more marked postorbital constriction and a larger nuchal crest; from *Ardipithecus*^{6,7} by upper I^1 with distinctive lingual topography characterized by extensive development of the crests and cingulum; less incisiform upper canines not diamond shaped with a low distal shoulder and a mesiodistal long axis, bucco-lingually narrower lower canines with stronger distal tubercle, and P_4 with two roots; from *Ororin*⁸ by upper I^1 with multiple tubercles on the lingual fossa, and non-chimp-like upper canines with extensive apical wear.

Type species. *Sahelanthropus tchadensis* sp. nov.

Etymology. In recognition that all specimens were recovered in Chad.
Holotype. TM 266-01-060-1, a nearly complete cranium with the following: on the right— I^2 alveolus, C (distal part), P^3 – P^4 roots, fragmentary M^1 and M^2 , M^3 ; and on the left— I^2 alveolus, C – P^4 roots, fragmentary M^1 – M^3 (Fig. 1 and Tables 1–5). Found by D.A. on 19 July 2001.

After study, the holotype and paratype series will be housed in the Département de Conservation des Collections, Centre National d'Appui à la Recherche (CNAR) in Ndjaména, Chad. The holotype has been dubbed 'Toumaï'; in the Goran language spoken in the Djurab Desert, this name is given to babies born just before the dry season, and means 'hope of life'.

Paratypes. See Table 1 for a list of paratypes, and Fig. 2 for illustrations.

Locality. Toros-Menalla locality TM 266, a single quarry of about 5,000 m², 16° 14' 30"–16° 15' 30" N, 17° 28' 30"–17° 30' 00" E (western Djurab Desert, northern Chad).

Horizon. All hominid specimens were found in the Toros-Menalla anthracotheriid unit (AU) and come from a perilacustrine sandstone¹². The associated fauna is biochronologically¹² older than fossils from Lukeino, Kenya (~6 Myr; refs 8, 9), and more closely resembles material from the Nawata formation at Lothagam, Kenya, which is radio-isotopically dated to 5.2–7.4 Myr (ref. 13). Biochronological studies are still underway and TM 266 fauna can be tentatively dated between 6 and 7 Myr (ref. 12).

Diagnosis. Same as for genus.

Preservation

All specimens are relatively well preserved, but almost the entire

Table 1 Specimens of *Sahelanthropus tchadensis* gen. et sp. nov.

Specimen number	Collected	Element	Discoverer	Dental dimensions (mm)
TM 266-01-060-1 (holotype) (Fig. 1)	2001	Cranium	D.A.	RC, BL = 10.2; RM ¹ , MD = 10.9; RM ² , MD = 13.0, BL = (12.8); RM ³ , MD = 10.8, BL = 14.9; LM ¹ , MD = 11.5
TM 266-01-060-2	2001	Symphyseal fragment with I and C alveoli	Group	RM ³ , MD = 10.7, BL = 12.7
TM 266-01-447	2001	Right M ³	Group	RI ¹ , MD = (13.3), BL = 8.9
TM 266-01-448 (Fig. 2a)	2001	Right I ¹	Group	RP ₄ , MD = 8.0;
TM 266-02-154-1 (Fig. 2b, c)	2002	Right mandible, (P_3) P_4 – M_3	D.A.	RM ₁ , MD = 11.0, BL = 11.9; RM ₂ , MD = 12.5; RM ₃ , MD = 13.3, BL = 12.2 Rc, MD = 11.0, BL = 8.5
TM 266-02-154-2 (Fig. 2d, e)	2002	Right c	D.A.	

Fossil hominids recovered from Toros-Menalla between July 2001 and February 2002. BL, buccolingual; L, left; MD, mesiodistal; R, right. Parentheses indicate estimated measurements—(P_3) indicates that only the roots are known.

cranium has been flattened dorsoventrally and the entire right side is depressed. The cranium exhibits broken but undistorted bone units and matrix-filled cracks (Fig. 1). Estimated measurements are given from a preliminary three-dimensional reconstruction. It will soon be possible to use computer tomography (CT) scans to generate a definitive three-dimensional reconstruction and stereo-lithographic casts.

Comparative observations

Cranial comparisons of comparably aged hominids are limited to a fragmentary basicranium and temporal bone: ARA-VP1/500 and ARA-VP1/125 (ref. 6) from 4.4 Myr *Ardipithecus ramidus ramidus*.

The next-oldest maxillary and cranial specimens are younger: KNM-KP-29283 (*A. anamensis*^{5,20}), KNM-WT40000 (*Kenyanthropus platyops*⁴), and *A. afarensis* specimens, which include the well-preserved AL 444-2 (ref. 21). The most notable anatomical features of the *S. tchadensis* cranium for comparative purposes are to be found in the face, which shows a mosaic of primitive and derived features. The face (Fig. 1a) is tall with a massive brow ridge, yet the mid-face is short (in the superoinferior dimension) and less prognathic than in *Pan* or *Australopithecus* (Table 3). This unusual combination of features is in turn associated with a relatively long braincase, comparable in size to those of extant apes (Fig. 1b, c). Preliminary comparisons with *Pan* suggest an endocranial volume

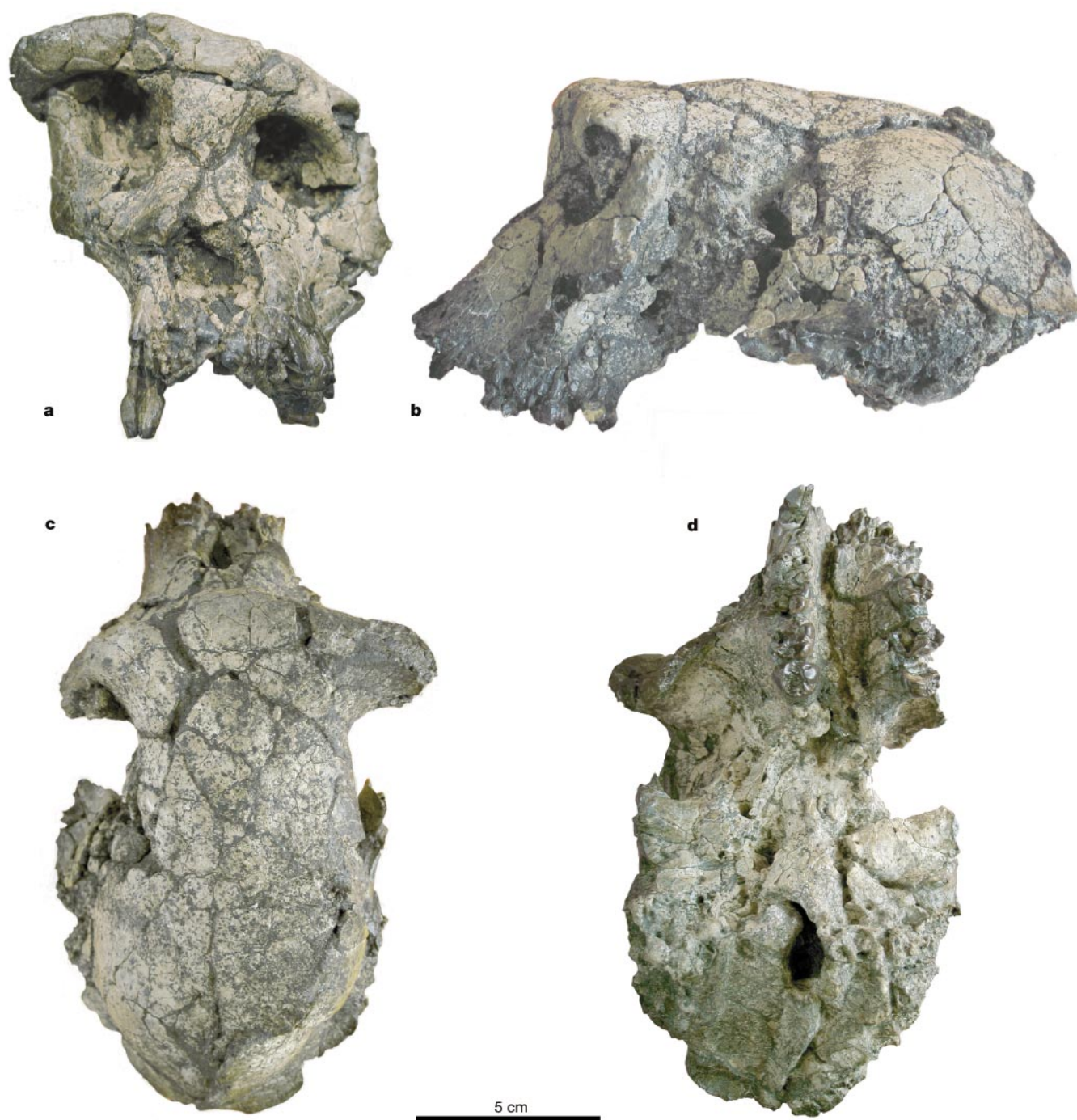


Figure 1 Cranium of *Sahelanthropus tchadensis* gen. et sp. nov. holotype (TM 266-01-060-1). **a**, Facial view. **b**, Lateral view. **c**, Dorsal view. **d**, Basal view.

of 320–380 cm³.

Although the *Sahelanthropus* cranium is considerably smaller than that of a modern male *Gorilla*, its supraorbital torus is relatively and absolutely thicker. This is probably a sexually dimorphic character (see Fig. 3), presumably reflecting strong sexual selection. If this is a male, then the combination of a massive

brow ridge with small canines suggests that canine size was probably not strongly sexually dimorphic. The interorbital pillar is wide (Fig. 1a). The zygomatic process of the maxilla emerges above the mesial margin of M¹ and is therefore more posterior relative to the cheek teeth than in *Australopithecus*^{19–21} and *Paranthropus*¹⁸. The infraorbital plane (from the lower orbital margin to the inferior

Table 2 Comparative dental measurements

		Mesiodistal					Buccolingual				
		<i>n</i>	Min.	Max.	Mean	s.d.	<i>n</i>	Min.	Max.	Mean	s.d.
Upper dentition											
I ¹	<i>S. tchadensis</i>	1	–	–	(13.3)	–	1	–	–	8.9	–
	<i>O. tugenensis</i> ⁸	1	–	–	(10.0)	–	1	–	–	8.7	–
	<i>A. r. ramidus</i> ⁶	1	–	–	(10.0)	–	2	7.5	8.2	–	–
	<i>A. anamensis</i> ²⁰	3	10.5	12.4	11.3	1.0	3	8.2	9.3	8.8	0.6
	<i>A. afarensis</i> ⁶	3	10.8	11.8	11.2	0.6	5	7.1	8.6	8.2	0.6
	<i>P. t. troglodytes</i> ²⁶	14	–	–	12.2	0.8	14	–	–	9.4	0.8
C	<i>S. tchadensis</i>	–	–	–	–	–	1	–	–	10.2	–
	<i>O. tugenensis</i> ⁸	1	–	–	11.0	–	1	–	–	9.3	–
	<i>A. r. ramidus</i> ⁶	2	(11.2)	11.5	–	–	2	11.1	11.7	–	–
	<i>A. anamensis</i> ²⁰	2	(10.6)	11.7	–	–	2	10.2	11.2	–	–
	<i>A. afarensis</i> ⁶	9	8.9	11.6	10.0	0.8	10	9.3	12.5	10.9	1.1
	<i>P. t. troglodytes</i> ²⁶	15	–	–	15.6	2.1	15	–	–	11.3	1.37
M ¹	<i>S. tchadensis</i>	2	(10.9)	(11.5)	–	–	–	–	–	–	–
	<i>A. r. kadabba</i> ⁷	1	–	–	(10.6)	–	1	–	–	12.1	–
	<i>A. anamensis</i> ²⁰	7	10.3	12.9	11.7	0.8	6	11.7	14.1	13.0	0.9
	<i>A. afarensis</i> ²⁵	14	10.5	13.8	12.2	1.0	12	12.0	15.0	13.4	0.9
	<i>P. t. troglodytes</i> ²⁶	14	–	–	10.5	0.5	14	–	–	11.3	0.6
M ²	<i>S. tchadensis</i>	1	–	–	13.0	–	1	–	–	(12.8)	–
	<i>A. r. ramidus</i> ⁶	2	(11.8)	11.8	–	–	2	(14.1)	(15.0)	–	–
	<i>A. anamensis</i> ²⁰	6	10.9	14.2	12.5	1.2	6	13.2	16.3	14.8	1.0
	<i>A. afarensis</i> ⁶	5	12.1	13.5	12.8	0.5	6	13.4	15.1	14.7	0.6
	<i>P. t. troglodytes</i> ²⁶	16	–	–	10.7	0.6	16	–	–	11.7	0.8
M ³	<i>S. tchadensis</i>	2	10.7	10.8	–	–	2	12.7	14.9	–	–
	<i>O. tugenensis</i> ⁸	2	10.2	10.3	–	–	2	12.9	13.1	–	–
	<i>A. r. kadabba</i> ⁷	1	–	–	10.9	–	1	–	–	12.2	–
	<i>A. r. ramidus</i> ⁶	1	–	–	10.2	–	1	–	–	12.3	–
	<i>A. anamensis</i> ²⁰	7	11.1	15.7	12.4	1.6	5	13.0	14.7	13.8	0.6
	<i>A. afarensis</i> ⁶	8	10.5	14.3	11.9	1.4	8	13.0	15.5	13.8	1.0
	<i>P. t. troglodytes</i> ²⁶	16	–	–	9.9	0.6	16	–	–	10.8	1.0
Lower dentition											
c	<i>S. tchadensis</i>	1	–	–	11.0	–	1	–	–	8.5	–
	<i>A. r. kadabba</i> ⁷	2	10.8	11.2	–	–	2	7.8	7.8	–	–
	<i>A. r. ramidus</i> ⁶	–	–	–	–	–	1	–	–	11.0	–
	<i>A. anamensis</i> ²⁰	7	6.6	10.4	9.0	1.3	6	9.2	11.4	10.2	1.0
	<i>A. afarensis</i> ²⁷	11	7.5	11.7	8.9	1.2	13	8.8	12.4	10.4	1.1
	<i>P. t. troglodytes</i> ²⁶	15	–	–	14.0	1.5	15	–	–	11.4	1.4
P ₄	<i>S. tchadensis</i>	1	–	–	8.0	–	–	–	–	–	–
	<i>O. tugenensis</i> ⁸	1	–	–	(8.0)	–	1	–	–	(9.0)	–
	<i>A. r. kadabba</i> ⁷	1	–	–	(8.1)	–	1	–	–	10.0	–
	<i>A. r. ramidus</i> ⁶	2	7.5	8.9	–	–	2	(9.9)	(11.5)	–	–
	<i>A. anamensis</i> ²⁰	8	7.4	9.8	8.8	1.0	9	9.6	11.9	10.7	0.8
	<i>A. afarensis</i> ⁶	15	7.7	11.1	9.7	1.0	14	9.8	12.8	10.9	0.8
	<i>P. t. troglodytes</i> ²⁶	15	–	–	8.1	0.6	16	–	–	8.8	0.8
M ₁	<i>S. tchadensis</i>	1	–	–	11.0	–	1	–	–	11.9	–
	<i>A. r. ramidus</i> ⁶	2	11.0	11.1	–	–	2	(10.2)	10.3	–	–
	<i>A. anamensis</i> ²⁰	11	11.5	13.8	12.6	0.9	12	10.5	14.8	12.1	1.3
	<i>A. afarensis</i> ⁶	17	11.2	14.0	13.0	0.6	16	11.0	13.9	12.6	0.8
	<i>P. t. troglodytes</i> ²⁶	15	–	–	10.7	0.4	15	–	–	9.2	0.6
M ₂	<i>S. tchadensis</i>	1	–	–	12.5	–	–	–	–	–	–
	<i>O. tugenensis</i> ⁸	1	–	–	(11.5)	–	1	–	–	(11.8)	–
	<i>A. r. kadabba</i> ⁷	1	–	–	(12.7)	–	1	–	–	11.8	–
	<i>A. r. ramidus</i> ⁶	1	–	–	(13.0)	–	1	–	–	11.9	–
	<i>A. anamensis</i> ²⁰	8	13.0	15.9	14.1	1.4	11	12.3	15.1	13.5	0.9
	<i>A. afarensis</i> ⁶	23	12.4	16.2	14.3	1.0	22	12.1	15.2	13.5	0.9
	<i>P. t. troglodytes</i> ²⁶	15	–	–	11.3	0.5	15	–	–	10.6	0.9
M ₃	<i>S. tchadensis</i>	1	–	–	13.3	–	1	–	–	12.2	–
	<i>O. tugenensis</i> ⁸	2	(12.3)	(12.4)	–	–	2	10.4	11.2	–	–
	<i>A. r. kadabba</i> ⁷	1	–	–	13.3	–	–	–	–	–	–
	<i>A. r. ramidus</i> ⁶	1	–	–	12.7	–	1	–	–	11.0	–
	<i>A. anamensis</i> ²⁰	6	13.7	17.0	14.6	1.2	6	11.9	13.4	12.8	0.7
	<i>A. afarensis</i> ⁶	14	13.7	16.3	14.8	0.8	14	12.1	14.9	13.3	0.8
	<i>P. t. troglodytes</i> ²⁶	16	–	–	11.0	0.6	16	–	–	9.6	0.8

Parentheses indicate estimated measurements.

Table 3 Hominid alveolar height measurements

Taxon	Alveolar height (mm)
<i>S. tchadensis</i> (TM 266-01-060-1)	(22)
<i>A. afarensis</i> ²¹ (AL 417-1, AL 444-2)	30, 33
<i>A. africanus</i> ²⁸ (mean and range, <i>n</i> = 11)	25.7 (21.1–30.0)
<i>P. boisei</i> ²⁸	42.2
<i>H. habilis</i> ²⁸ (KNM-ER 1470, KNM-ER 1813)	31.0, 25.0

Alveolar height from nasospinale to prosthion. Parentheses indicate estimated measurements from a preliminary three-dimensional reconstruction.

malar margin) is similar to that of *Pan* and differs from both *Gorilla* and larger *Australopithecus*, in which this region is absolutely and relatively taller. The canine fossa is similar to that in AL 444-2 (ref. 21) and there is no diastema between the alveoli of I² and C in the Chadian specimen. There is no supratoral sulcus. Between the temporal lines, the frontal squama is slightly depressed but not like the frontal trigone usually seen in *P. boisei*¹⁸. The temporal lines converge behind the coronal suture as in crested *Pan*, whereas their junction is more anterior than generally seen in some large *A. afarensis* (for example, AL 444-2; ref. 21). The sagittal crest is a little larger posteriorly than in either AL 444-2 (ref. 21) or KNM-WT 40000 (ref. 4). The compound temporal–nuchal crest is marked as in KNM-ER 1805 (ref. 22) and much larger than in the male *A. afarensis* AL 444-2 (ref. 21), suggesting a relatively large posterior temporalis muscle. The postorbital breadth (Fig. 1c) is absolutely smaller than the *Pan*–*Gorilla* average, and similar to that in AL 444-2 and AL 417-1 (ref. 21) (Table 5).

The basicranium (Fig. 1d) has small occipital condyles associated with an apparently large foramen magnum. Despite damage, the

Table 4 Cranio-facial measurements of *S. tchadensis*

Cranio-facial features	Size (mm)
Orbital height × width	(36 × 35)
Maximum breadth	
Bicarotid chord	(45)
Bimastoid (at the nuchal crest level)	(108)
Biporion	(102)
Mastoid mesiodistal length	54.5
Nuchal plane length (opisthion–inion)	(47)

Measurements are from TM 266-01-060-1. Parentheses indicate estimated measurements from a preliminary three-dimensional reconstruction.

foramen magnum seems to be longer than wide, and not like the rounded shape typical of *Pan*. As in *A. ramidus*, the basion is intersected by the bicarotid chord; the basion is posterior in large apes and anterior in some of the later hominids. The *Sahelanthropus* basioccipital is correlatively short and shaped like a truncated triangle as in *Ardipithecus*⁶; it is not as triangular as in other early hominids.

Like *Pan*, *Australopithecus*²¹ and *Ardipithecus*⁶, the orientation of the petrous portion of the temporal is approximately 60° relative to the bicarotid chord, instead of 45° as is typical of *Paranthropus* and *Homo*. Unlike *Gorilla*, *Pan* and *Ardipithecus*⁶, the posterior margin of the tympanic tube does not have a crest-like morphology. Compared with known *Ardipithecus* (ARA-VP-1/500)⁶, the glenoid cavity is larger and the postglenoid process mediolaterally wider, so the pronounced squamotympanic fissure is situated more medially to the postglenoid process. The height of the relatively flat temporomandibular joint above the tooth row suggests a high ascending

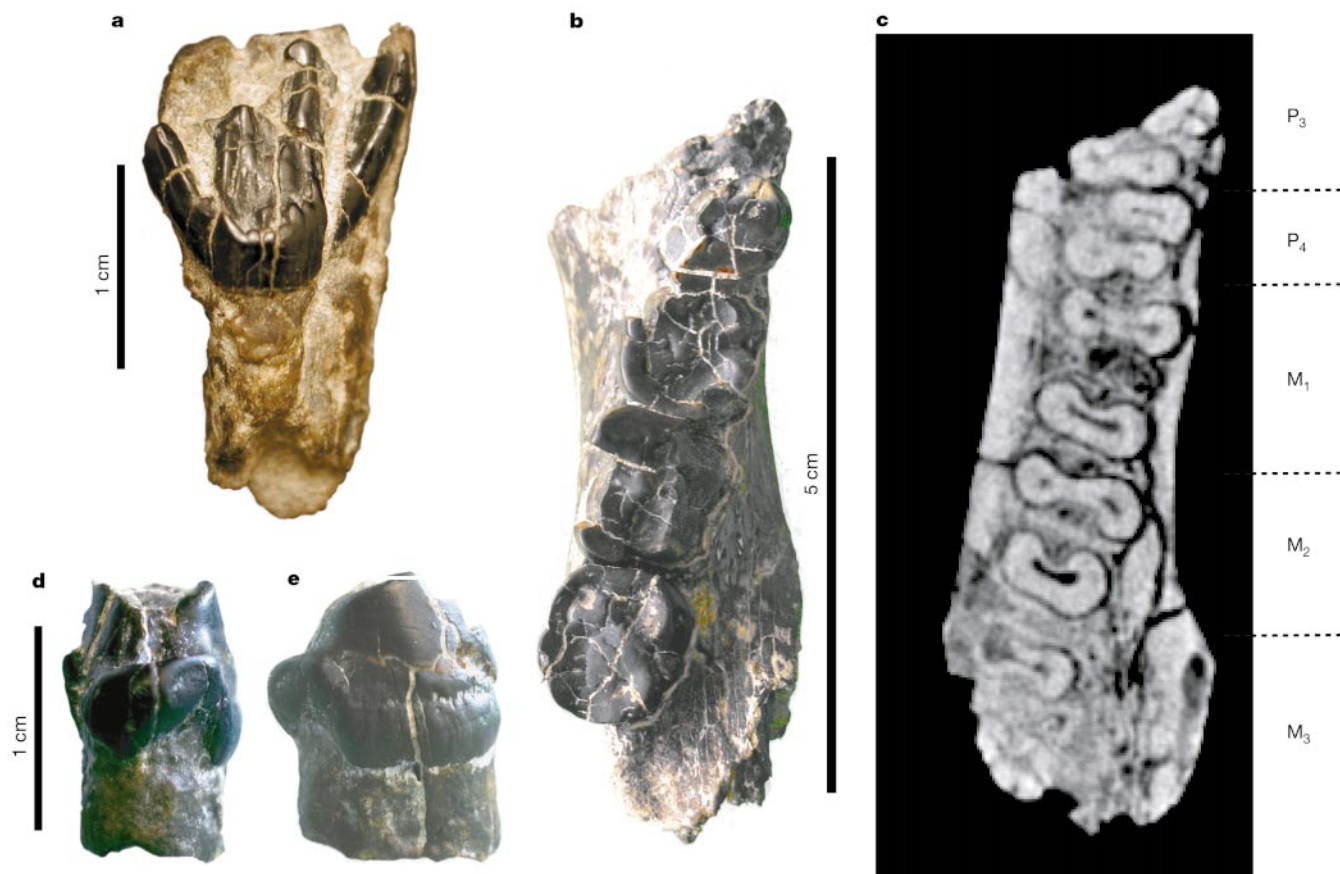


Figure 2 *Sahelanthropus tchadensis* gen. et sp. nov. paratypes. **a**, Right upper I¹, lingual view (TM 266-01-448). **b**, **c**, Right lower jaw (TM 266-02-154-1), occlusal view (**b**) and axial CT scan (**c**). **d**, **e**, Right lower canine (TM 266-02-154-2), distal view (**d**) and buccal view (**e**).

Table 5 Measurements and indices of hominoid frontal bones

	Post orbital breadth* (a) (mm)	Superior facial breadth† (b) (mm)	Fronto-facial breadth index (a/b × 100)
<i>P. troglodytes</i> (10 males) ¹⁹	71.0 (62.0–76.5)	110.1 (99.4–129.0)	64.4 (58.1–68.5)
<i>P. troglodytes</i> (10 females) ¹⁹	70.1 (65.0–76.0)	102.2 (87.3–112.0)	68.8 (66.6–74.5)
<i>G. gorilla</i> (10 males) ¹⁹	69.5 (60.5–77.0)	135.1 (127.7–144.0)	51.5 (47.4–59.4)
<i>G. gorilla</i> (10 females) ¹⁹	68.5 (65.5–73.5)	115.5 (108.0–127.0)	59.5 (57.0–63.4)
<i>P. pygmaeus</i> (5 males) ¹⁹	63.8 (59.0–69.5)	105.7 (88.3–116.2)	61.2 (51.7–72.5)
<i>P. pygmaeus</i> (4 females) ¹⁹	64.0 (61.5–65.5)	90.1 (87.8–94.0)	71.1 (68.1–74.0)
<i>S. tchadensis</i> gen. et sp. nov. (TM 266-01-060-1)	(60)	(102)	(59)
<i>A. afarensis</i> (AL 444-2) ²¹	77.0	118.6	64.9

Values are mean and range. Values for *S. tchadensis* were estimated from a preliminary three-dimensional reconstruction.

*The least-frontal breadth.

†Left frontomale temporale to right frontomale temporale (outer biorbital breadth).

ramus of the mandible more similar in absolute size to *Gorilla* than to *Pan*. The large pneumatized mastoid process is anteroposteriorly long, more so than in *Ardipithecus*⁶ and *Australopithecus* (AL 444-2, AL 333)^{19,21}. The flat nuchal plane is relatively longer (Fig. 1b, d and Table 4) than in *Pan*, *Gorilla*, AL 444-2 and AL 333, and with crests as marked as those of *Gorilla*, implying the presence of relatively large superficial neck muscles. The horizontally oriented nuchal plane is much flatter than the convex nuchal plane of *Pan*. There is not yet sufficient information to infer reliably whether *Sahelanthropus* was a habitual biped. However, such an inference would not be unreasonable given the skull's other basicranial and facial similarities to later fossil hominids that were clearly bipedal.

Little is preserved of the mandibular corpus in TM-266-02-154-1 or the symphysis in TM-266-01-060-2 to permit much discussion of mandibular anatomy. However, the former, presumably a male, has a thick corpus with a wide extramolar sulcus (Fig. 2a).

In dentition, both lower incisor roots are small. The lingual surface of upper I¹ displays distinct marginal ridges converging basally onto a narrow gingival eminence and with several tubercles extending incisally into the lingual fossa (Fig. 2a). The smaller upper I² alveolus is situated, in frontal view, lateral to the lateral edge of the nasal aperture (Fig. 1a), a probable symplesiomorphy with *Pan*. The upper and lower canines are small (Tables 1, 2). Given the absolutely and relatively massive supraorbital torus of the cranium (Figs 1a and 3), possibly reflecting strong sexual selection, and the thick corpus of the mandible (Fig. 2b, c), which probably indicate male status, we infer that *Sahelanthropus* canines were probably weakly sexually dimorphic. The upper canine (Fig. 1d) has both distal and apical wear facets whereas M³ is unworn. The lower canine (Fig. 2d, e) has a strong distal tubercle that is separated from a distolingual crest by a fovea-like groove; the large apical wear zone at a level above this distal tubercle implies a non-honing C–P₃ complex (the

P₃ is still unknown). The upper canine, judging from the steep, narrow distal wear strip reaching basally, we believe had a somewhat lower distal shoulder than *Ardipithecus*^{6,7}, suggesting an earlier evolutionary stage. Moreover, a small, elliptical contact facet for P₃ on the distobuccal face of the distal tubercle indicates the absence of a lower c–P₃ diastema. *Sahelanthropus* thus probably represents an early stage in the evolution of the non-honing C–P₃ complex characteristic of the later hominids⁷.

The cheek teeth are small (Tables 1, 2), within the size range of *A. ramidus* and the lower end of *A. afarensis*. Lower c and the lower and upper premolars each have three pulp canals and two roots (Fig. 2c). The P₃ has a 2R:MB + D pattern (terminology following ref. 22: R, root; M, mesial; MB, mesiobuccal; D, distal) with a small mesio-buccal root and two partially fused, larger oblique distal roots. The P₄ has two transverse roots (2R:M + D), with the mesial one smaller than the distal; in both premolars the distal root has two pulp canals (Fig. 2c). The P₄ teeth of *Ardipithecus* have more-derived root patterns, with either a Tome's root (*A. r. kadabba*)⁷ or a single root (*A. r. ramidus*)⁶. *S. tchadensis* has a P₄ with a large talonid and well-marked grooves on the buccal face. The molar size gradient is M₂ > M₃ > M₁. The upper molars have three roots, two buccal and one lingual, which is large and mesiodistally elongated. The M³ crowns are triangular whereas the M₃ teeth are rectangular and rounded distally. The lower molar root pattern is 2R:M + D: a mesial root with two pulp canals and a distal root with only one canal (Fig. 2c).

The radial enamel thickness is 1.71 mm at paracone LM³ and 1.79 mm at hypocone LM². The molar enamel is therefore thicker than in *Pan*, possibly thicker than in *Ardipithecus ramidus ramidus*, but thinner than in *Australopithecus*⁶. Given individual and intra-specific variation, further study using high-resolution imaging (CT scanning) is needed for useful comparisons⁷.

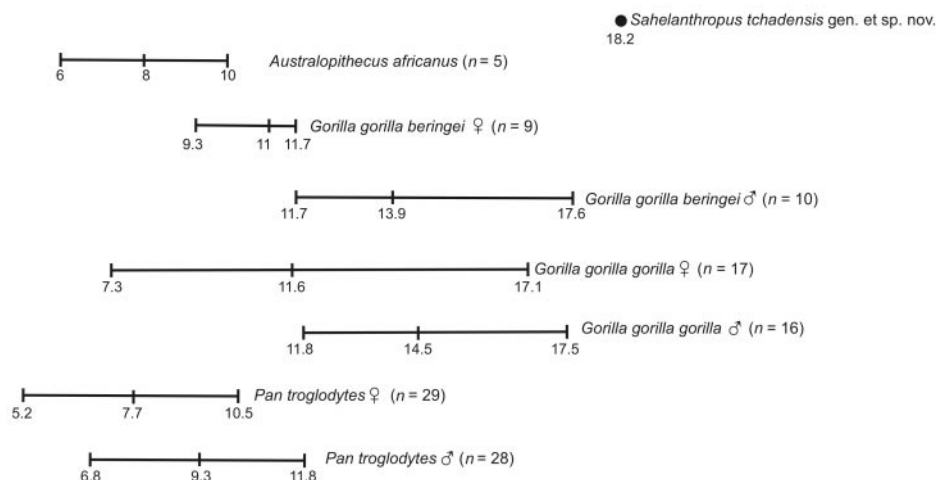


Figure 3 Vertical (superoinferior) thickness of the supraorbital torus in extant hominoids, *A. africanus* and *S. tchadensis* gen. et sp. nov. Measurements are in millimetres at

the highest point of the superior orbital margin. Data for apes and *A. africanus* are from ref. 28.

Discussion

Sahelanthropus has several derived hominid features, including small, apically worn canines—which indicate a probable non-honing C–P₃ complex—and intermediate postcanine enamel thickness. Several aspects of the basicranium (length, horizontal orientation, anterior position of the foramen magnum) and face (markedly reduced subnasal prognathism with no canine diastema, large continuous supraorbital torus) are similar to later hominids including *Kenyanthropus* and *Homo*. All these anatomical features indicate that *Sahelanthropus* belongs to the hominid clade.

In many other respects, however, *Sahelanthropus* exhibits a suite of primitive features including small brain size, a truncated triangular basioccipital bone, and the petrous portion of the temporal bone oriented 60° to the bicarotid chord. The observed mosaic of primitive and derived characters evident in *Sahelanthropus* indicates its phylogenetic position as a hominid close to the last common ancestor of humans and chimpanzees. Given the biochronological age of *Sahelanthropus*, the divergence of the chimpanzee and human lineages must have occurred before 6 Myr, which is earlier than suggested by some authors^{23,24}. It is not yet possible to discern phylogenetic relationships between *Sahelanthropus* and Upper Miocene hominoids outside the hominid clade. *Ouranopithecus*¹⁵ (about 2 Myr older) is substantially larger, with quadrate orbits, a very prognathic and wide lower face, large male canines with a long buccolingual axis, and cheek teeth with very thick enamel. *Samburupithecus*¹⁴ (about 2.5 Myr older) has a low, posteriorly positioned (above M²) zygomatic process of the maxilla, cheek teeth with high cusps (similar to *Gorilla*), lingual cingula, large premolars and a large M³.

Sahelanthropus is the oldest and most primitive known member of the hominid clade, close to the divergence of hominids and chimpanzees. Further analysis will be necessary to make reliable inferences about the phylogenetic position of *Sahelanthropus* relative to known hominids. One possibility is that *Sahelanthropus* is a sister group of more recent hominids, including *Ardipithecus*. For the moment, productive comparisons of *Sahelanthropus* with *Orrorin* are difficult because described craniodental material of the latter is fragmentary and no *Sahelanthropus* postcrania are available. However, we note that in *Orrorin*, the upper canine resembles that of a female chimpanzee. The discoveries of *Sahelanthropus* along with *Ardipithecus*^{6,7} and *Orrorin*⁸ indicate that early hominids in the late Miocene were geographically more widespread than previously thought.

Finally, we note that *S. tchadensis*, the most primitive hominid, is from Chad, 2,500 km west of the East African Rift Valley. This suggests that an exclusively East African origin of the hominid clade is unlikely to be correct (contrary to ref. 8). It will never be possible to know precisely where or when the first hominid species originated, but we do know that hominids had dispersed throughout the Sahel and East Africa¹⁰ by 6 Myr. The recent acquisitions of Late Miocene hominid remains from three localities, as well as functional, phylogenetic and palaeoenvironmental studies now underway, promise to illuminate the earliest chapter in human evolutionary history. *Sahelanthropus* will be central in this endeavour, but more surprises can be expected. □

Received 13 March; accepted 27 May 2002; doi:10.1038/nature00879.

1. Dart, R. A. *Australopithecus africanus*, the man-ape of South Africa. *Nature* **115**, 195–199 (1925).
2. Kortlandt, A. *New Perspectives on Ape and Human Evolution* 9–100 (Stichting voor Psychobiologie, Amsterdam, 1972).
3. Coppens, Y. *Le singe, l'Afrique et l'Homme* (Fayard, Paris, 1983).
4. Leakey, M. G. *et al.* New hominid genus from eastern Africa shows diverse middle Pliocene lineages. *Nature* **410**, 433–440 (2001).

5. Leakey, M. G., Feibel, C. S., McDougall, I. & Walker, A. C. New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. *Nature* **376**, 565–571 (1995).
6. White, T. D., Suwa, G. & Asfaw, B. *Australopithecus ramidus*, a new species of hominid from Aramis, Ethiopia. *Nature* **371**, 306–312 (1994).
7. Haile-Selassie, Y. Late Miocene hominids from the Middle Awash, Ethiopia. *Nature* **412**, 178–181 (2001).
8. Senut, B. *et al.* First hominid from the Miocene (Lukeino Formation, Kenya). *C. R. Acad. Sci. Paris* **332**, 137–144 (2001).
9. Deino, A. L., Tauxe, L., Monaghan, M. & Hill, A. ⁴⁰Ar/³⁹Ar geochronology and paleomagnetic stratigraphy of the Lukeino and lower Chemeron Formations at Tabarin and Kapcheberek, Tugen Hills, Kenya. *J. Hum. Evol.* **42**, 117–140 (2002).
10. Brunet, M. *et al.* The first australopithecine 2500 kilometres west of the Rift Valley (Chad). *Nature* **378**, 273–275 (1995).
11. Brunet, M. *et al.* *Australopithecus bahrelghazali*, une nouvelle espèce d'Hominidé ancien de la région de Koro Toro (Tchad). *C. R. Acad. Sci. Paris* **322**, 907–913 (1996).
12. Vignaud, P. *et al.* Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* **418**, 152–155 (2002).
13. McDougall, I. & Feibel, C. S. Numerical age control for the Miocene–Pliocene succession at Lothagam, a hominoid-bearing sequence in the northern Kenya Rift. *J. Geol. Soc. Lond.* **156**, 731–745 (1999).
14. Ishida, H. & Pickford, M. A new Late Miocene hominoid from Kenya: *Samburupithecus kiptalami* gen. et sp. nov. *C. R. Acad. Sci. Paris* **325**, 823–829 (1998).
15. Bonis, L. de & Koufos, G. D. The face and the mandible of *Ouranopithecus macedoniensis*: description of new specimens and comparisons. *J. Hum. Evol.* **24**, 469–491 (1993).
16. Pilbeam, D. New hominoid skull material from the Miocene of Pakistan. *Nature* **295**, 232–234 (1982).
17. Kordos, L. & Begun, D. R. A new cranium of *Dryopithecus* from Rudabanya, Hungary. *J. Hum. Evol.* **41**, 689–700 (2001).
18. Tobias, P. V. *The Cranium and Maxillary Dentition of Australopithecus (Zinjanthropus) boisei*, Olduvai Gorge (Cambridge Univ. Press, London, 1967).
19. Kimbel, W. H., White, T. D. & Johanson, D. C. Cranial morphology of *Australopithecus afarensis*: A comparative study based on a composite reconstruction of the adult skull. *Am. J. Phys. Anthropol.* **64**, 337–388 (1984).
20. Ward, C. V., Leakey, M. G. & Walker, A. Morphology of *Australopithecus anamensis* from Kanapoi and Allia Bay, Kenya. *J. Hum. Evol.* **41**, 235–368 (2001).
21. Kimbel, W. H., Johanson, D. C. & Rak, Y. The first skull and other new discoveries of *Australopithecus afarensis* at Hadar, Ethiopia. *Nature* **368**, 449–451 (1994).
22. Wood, B. *Koobi Fora Research Project: Hominid Cranial Remains* Vol. 4 (Clarendon, Oxford, 1991).
23. Kumar, S. & Hedges, B. A molecular time scale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
24. Pilbeam, D. in *The Primate Fossil Record* (ed. Hartwig, W.) 303–310 (Columbia Univ. Press, New York, 2002).
25. Kimbel, W. H. Systematic assessment of a maxilla of *Homo* from Hadar, Ethiopia. *Am. J. Phys. Anthropol.* **103**, 235–262 (1997).
26. Uchida, A. *Intra-species Variation among the Great-Apes: Implications for Taxonomy of Fossil Hominoids* Thesis, Harvard Univ. (1992).
27. Lockwood, C. A., Kimbel, W. H. & Johanson, D. C. Temporal trends and metric variation in the mandibles and dentition of *A. afarensis*. *J. Hum. Evol.* **39**, 23–55 (2000).
28. Lockwood, C. A. Sexual dimorphism in the face of *Australopithecus africanus*. *Am. J. Phys. Anthropol.* **108**, 97–127 (1999).

Acknowledgements

We thank the Chadian Authorities (Ministère de l'Éducation Nationale de l'Enseignement Supérieur et de la Recherche, Université de N'djaména, CNAR). We extend gratitude for their support to the French Ministries, Ministère Français de l'Éducation Nationale (Faculté des Sciences, Université de Poitiers), Ministère de la Recherche (CNRS), Ministère des Affaires Étrangères (Direction de la Coopération Scientifique, Universitaire et de Recherche, Paris, and SCAC Ambassade de France à N'djaména), to the Région Poitou-Charentes, le Département de la Vienne, the Association pour le Prix scientifique Philip Morris, and also to the Armée Française (MAM and Epervier) for logistic support. For giving us the opportunity to work with their collections, we are grateful to the National Museum of Ethiopia, the National Museum of Kenya, the Peabody Museum and Harvard University, the Institute of Human Origins and the University of California. Special thanks to Scanner-IRM Poitou Charentes (P. Chartier and F. Perrin), for industrial scanner to EMPA (A. Flisch Ing. HTL) and to Multimedia Laboratorium-Computer Department, University of Zurich-Irchel (P. Stucki). Many thanks to all our colleagues and friends for their help and discussion, and particularly to F. Clark Howell, A. Garaudel, Y. Haile-Selassie, D. Johanson, W. Kimbel, M. G. Leakey, D. Lieberman, R. Macchiarelli, M. Pickford, B. Senut, G. Suwa, T. White and Lubaka. We especially thank all the other MPFT members who joined us for the field missions, and V. Bellefret, S. Riffaut and J.-C. Bertrand for technical support. We are most grateful to G. Florent for administrative guidance. We dedicate this article to J. D. Clark. All authors are members of the MPFT.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.B. (e-mail: michel.brunet@univ-poitiers.fr).

Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad

Patrick Vignaud*, Philippe Dourigner†, Hassane Taïssou Mackaye‡, Andossa Likius*, Cécile Blondel*, Jean-Renaud Boissérie*, Louis de Bonis*, Véra Eisenmann§, Marie-Esther Etienne*, Denis Geraads||, Franck Guy¶, Thomas Lehmann*, Fabrice Lihoreau*, Nieves Lopez-Martinez#, Cécile Mourer-Chauviré☆, Olga Otero*, Jean-Claude Rage§, Mathieu Schuster†, Laurent Viriot*, Antoine Zazzo** & Michel Brunet*

* Faculté des Sciences et Centre National de Recherche Scientifique UMR 6046, Université de Poitiers, 40 Avenue du Recteur Pineau, 86022 Poitiers Cedex, France

† Centre de Géochimie de la Surface, CNRS UMR 7517, Université Louis Pasteur, 1 rue Blessig, 67084 Strasbourg, France

‡ Université de N'Djaména, BP 1117 N'Djaména, Tchad

§ Muséum National d'Histoire Naturelle et CNRS UMR 8569, rue Cuvier, 75005 Paris, France

|| Centre National de Recherche Scientifique UPR 2147, 44 rue de l'Amiral Mouchez, 75014 Paris, France

¶ Peabody Museum, Harvard University, 11 Divinity Avenue, Cambridge, Massachusetts, 02138, USA

Dept. Paleontología, Universidad Complutense, 28040 Madrid, Spain

☆ Centre des Sciences de la Terre, CNRS UMR 5125, Université Claude Bernard, 27-43 Bd du 11 novembre 1918, 69622 Villeurbanne, France

** Centre National de Recherche Scientifique UMR 162 et Institut National de la Recherche Agronomique, Université Pierre et Marie Curie, 4 place Jussieu, 75252 Paris Cedex 05, France

All six known specimens of the early hominid *Sahelanthropus tchadensis* come from Toros-Menalla site 266 (TM 266), a single locality in the Djurab Desert, northern Chad, central Africa. Here we present a preliminary analysis of the palaeontological and palaeoecological context of these finds. The rich fauna from TM 266 includes a significant aquatic component such as fish, crocodiles and amphibious mammals, alongside animals associated with gallery forest and savannah, such as primates, rodents, elephants, equids and bovids. The fauna suggests a biochronological age between 6 and 7 million years. Taken together with the sedimentological evidence, the fauna suggests that *S. tchadensis* lived close to a lake, but not far from a sandy desert, perhaps the oldest record of desert conditions in the Neogene of northern central Africa.

Since 1994, the Mission Paléoanthropologique Franco-Tchadienne (MPFT), a scientific collaboration between Poitiers University, Ndjamená University and the Centre National d'Appui à la Recherche (CNAR), Ndjamená, has conducted several field expeditions in the Djurab Desert of northern Chad (Fig. 1). Numerous Miocene and Pliocene vertebrate sites have been discovered, corresponding to four fossiliferous areas, all dated biochronologically—that is, by the evolutive degree of their faunas. These areas are Koro Toro, estimated at around 3.0–3.5 million years (Myr), which has already yielded hominid remains^{1,2}; Kollé, between 4 and 5 Myr (ref. 3); and Kossom Bougoudi, dated close to the Mio-Pliocene boundary around 5.3 Myr (ref. 4). The fourth fossiliferous area, named Toros-Menalla, is the oldest: it was discovered in the western Djurab Desert by the MPFT in 1997. More than 300 vertebrate localities are already known from this area. One of them, TM 266, has yielded a vertebrate fauna (42 taxa) that includes more than 700 identifiable mammal fossils (24 taxa) and the oldest known hominid remains⁵.

The Toros-Menalla fossiliferous area is located in the intracratonic Chad Basin (Fig. 1). This basin comprises a southern sub-basin that includes the present Lake Chad under semi-arid to wet conditions, and a northern sub-basin that forms the central Chad lowland, now a desert and subject to significant aeolian deflation. The Holocene 'Mega Lake Chad'⁶ constitutes the last main lacustrine event in this region before aeolian erosion and the onset of the current desert conditions. Sediments in the Chad Basin are mainly lacustrine, fluvial and aeolian. Fluvial deposits are mainly due to episodic flash floods. Until now, no perennial fluvial systems have been recognized in Miocene deposits of the northern sub-basin.

Sedimentology

The Toros-Menalla outcrops consist of large flat to weakly undulat-

ing surfaces, sometimes dissected in small outliers. Stratigraphic sections are rarely thicker than a few metres. Outcrops are situated in the centre of the northern sub-basin and consist mainly of thick monotonous terrigenous series dominated by sands and weakly indurated sandstones interbedded with argillaceous pelites and diatomites (Fig. 2).

The section and the depositional facies described here (Fig. 2) correspond to the TM 266 hominid locality (1.5 km², hominids quarry about 5,000 m²). Numerous sections (2–5 m thick) were examined and widely correlated. Having only small outliers, most of the sections were completed by excavations.

The lower part of the section (at least 4 m thick) is composed of fine to very fine white sands, poorly cemented, and is mainly constituted by numerous quartz grains, without matrix. The grains are well sorted, well rounded, matt and frosted, and are strong evidence for aeolian modelling. The foreset laminations (avalanche laminations in front of the aeolian dune) represent a typically aeolian deposit. These sands show cross-beddings that progressively decrease in size from the bottom (1–2 m) to the top (20 cm). This facies exhibits typical alternations of grain-fall and grain-flow laminations, characteristic of aeolian dune deposits^{7,8}. Our interpretation is confirmed by frequent wind ripples at the foot of the fossil dunes, whose crests are perpendicular to the direction of dune progradation. These fossil dunes are, to our knowledge, the oldest evidence for desert conditions in the southern Sahara area. The measured currents of the larger-scale cross-bedding of the fossil dunes show a unimodal distribution that indicates a major wind direction towards the west-south-west (Fig. 2). This facies contains numerous rhizoliths (rocks formed from plant roots), their number increasing gradually from the bottom to the top. In contrast to the others, this unit contains no vertebrate fauna. The erosional surface at its top marks the end of the true desertic episode.

The middle part of the section (about 2 m thick) is informally named the anthracotheriid unit (AU). This unit yielded the hominid cranium and all the terrestrial vertebrate remains (Table 1). The hominid was embedded in a poorly cemented sandstone in the lowest metre of the unit (Fig. 2). The deposit is a moderately to well-cemented sandstone, generally well sorted but with a variable quartz grain maturity. The sandstone is characterized by a mixing of both inherited aeolian sand showing matt and frosted quartz grains, and other grains resulting from the lacustrine reworking. In contrast to the lower part of the section, these sandstones are often matrix supported (mainly pelites and diatomites). There are many millimetre- to centimetre-scale lacustrine joints that contain clay. In contrast to the lower part of the section, the cross-bedding sets are always small (generally 5–10 cm, exceptionally 20 cm). This root-rich unit yielded almost all the vertebrate remains, common fossilized dung beetle brood balls and rare termitaries^{9,10}. In contrast to the lower aeolian facies, the occurrence of both current and wave ripples, as well as small sand bars of coarse pelitic sandstone, testify to water movements. However, in this unit the strong dispersion of the palaeocurrents in all directions (360°) (Fig. 2) is incompatible with a normal fluvial dynamic, in which currents never show such unusual dispersion (even in strongly meandering systems). This water circulation between the dunes represents the first manifestation of a northward lake transgression, which implicates episodic flooding and draining in a desert environment. The AU corresponds to a shallow perilacustrine environment, subject to frequent inundation due to recurrent lateral variations of the shoreline. For the present-day Lake Chad, such excursions may reach several tens of kilometres over one or two decades¹¹. In this vegetated perilacustrine belt, numerous environments coexist, from aquatic habitats in

the lake and along the shore, to very dry situations on the edge of the desert. The wide variation in palaeocurrent direction suggests the interdigitation and coexistence of lacustrine, perilacustrine and desert environments.

The upper part of the section (about 0.5 m thick) is weakly developed in the TM 266 site. It meets the middle part of the section at an erosional surface. It is a well-stratified green pelite and is interpreted as a true lacustrine environment. This upper unit yielded only aquatic vertebrates (fish and crocodilians).

Vertebrate fauna

The fauna recovered at TM 266 (Table 1) does not show evidence of fluvial transport, consistent with the sedimentology.

There are more than ten taxa of freshwater fish. All identified taxa are known to have been present in the Nilo-Sudanese at least since the Upper Miocene; all have living representatives in Lake Chad except *Sindacharax*, which is a taxon from the Miocene and Pliocene of north equatorial Africa¹². Turtles, lizards, snakes and crocodiles are represented by abundant skeletal material, but no bird remains have been recovered.

The carnivore fauna is particularly diverse. As in most Late

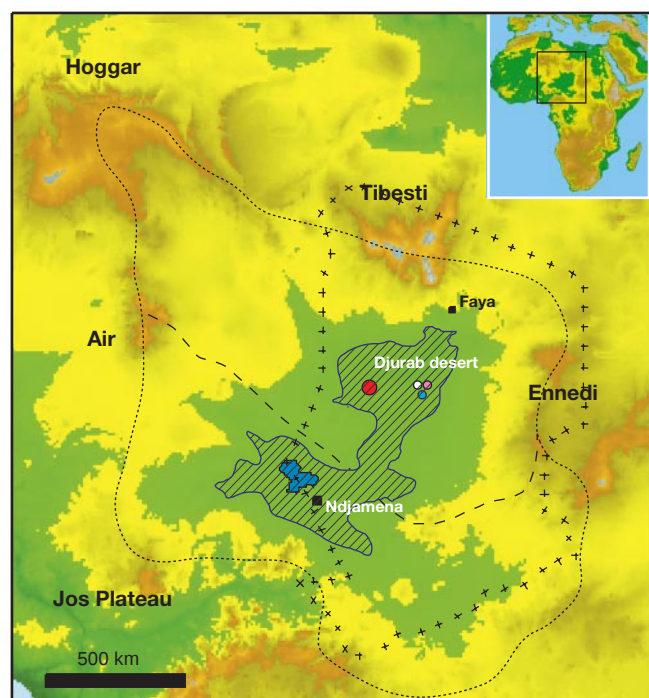


Figure 1 Map of the Chad basin showing the location of the Toros-Menalla area. The crossed line represents the Chad border. The boundaries of the ancient Lake Chad basin are shown as a dotted line (with northern and southern sub-basins divided by a dashed line). The inferred maximum extension of the Holocene 'Lake Mega Chad' is hatched in black; the small blue area within it corresponds to present-day Lake Chad. Red circle, the Toros-Menalla hominid site (TM 266); white circle, the Kossom-Bougoudi area; pink circle, Kollé area; blue circle, Koro-Toro area. Colours represent altitude: green, 280–430 m; yellow to grey, 430–850 m; brown, 850–2,000 m; grey-blue patches in Tibesti area, about 2,000–3,400 m.

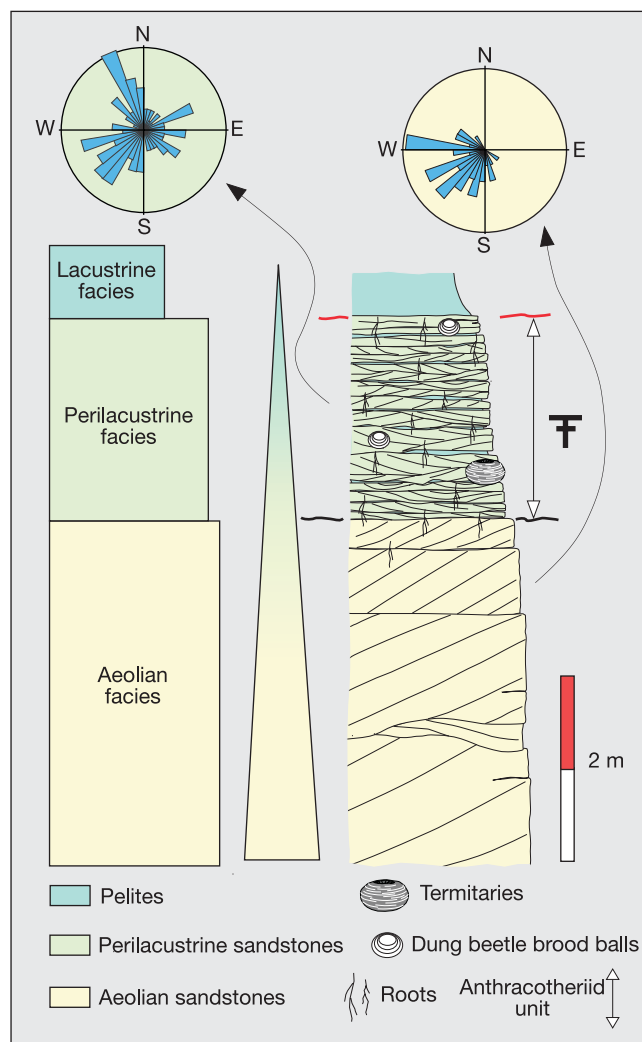


Figure 2 Idealized section and palaeoenvironmental vertical evolution in the TM 266 hominid site. Note the unimodal distribution of the direction of the currents measured in the aeolian facies (lower part of the section) in contrast to the variety of directions of the currents in the perilacustrine facies (middle part of section, anthracotheriid unit). The bottom of the crossed T indicates the precise location of the hominid remains.

Table 1 Flora and fauna of the TM 266 hominid locality

FLORA	MAMMALIA
LEGUMINOSAE	Carnivora
(?) Papilionoideae indet.	Hyaenidae
	<i>Hyaenictitherium</i> cf. <i>H. hyaenoides</i>
ARTHROPODA	<i>Ictitherium</i> sp. (size of <i>I. taurinum</i>)
COLEOPTERA	Hyaenidae indet. (size of <i>Protictitherium</i>)
Scarabaeidae indet. (nests)	Felidae
ISOPTERA	<i>Machairodus</i> cf. <i>M. giganteus</i>
Hodotermitidae indet. (nests)	Mustelidae
	Lutrinae indet.
VERTEBRATA	Primates
TELEOSTEI	Cercopithecidae
Momyriiformes	Colobinae indet.
Gymnarchidae	Hominidae
<i>Gymnarchus</i> sp.	Hominidae gen. et sp. nov.
Cypriniformes	Rodentia
Cyprinidae	Sciuridae
<i>Labeo</i> sp.	<i>Xerus</i> sp.
Characiformes	Muridae
Characidae	Murinae indet.
<i>Hydrocynus</i> sp.	Hystricidae
<i>Sindacharax</i> sp.	<i>Hystrix</i> sp.
<i>Alestini</i> indet.	Tubulidentata
Siluriformes	Orycteropodidae
Bagridae	<i>Orycteropus</i> cf. <i>O. gaudryi</i>
<i>Bagrus</i> group	Perissodactyla
Siluriformes indet.	Equidae
Perciformes	<i>Hipparion</i> cf. <i>H. abudhabiense</i>
Perciformes indet.	Proboscidea
Tetraodontiformes	Gomphotheriidae
Tetraodontidae	<i>Anancus kenyensis</i>
<i>Tetraodon</i> sp.	Elephantidae
Polypteriformes	<i>Loxodonta</i> sp. aff. <i>L. sp. indet.</i> "Lukeino stage"
Polypteridae	Artiodactyla
<i>Polypterus</i> sp.	Suidae
	<i>Nyanzachoerus syrticus</i>
REPTILIA	Anthracotheiidae
Testudines	<i>Libycosaurus petrocchii</i>
Trionychidae	Hippopotamidae
Trionychidae indet.	<i>Hexaprotodon</i> sp. nov.
Testudinidae	Giraffidae
Testudinidae indet.	<i>Sivatherium</i> cf. <i>S. hendeyi</i>
Serpentes	Giraffidae indet.
Boidae	Bovidae
<i>Python</i> cf. <i>P. sebae</i>	Antilopinae
Colubridae	Antilopini indet.
Colubridae indet.	Caprinae
Lacertilia	Ovibovini indet. aff. <i>Palaeoryx</i>
Varanidae	Bovinae
<i>Varanus</i> sp.	Bovini indet.
Crocodylia	Hippotraginae
Crocodylidae	Reduncini
<i>Crocodylus niloticus</i>	<i>Kobus</i> sp.
<i>Euthecodon</i> cf. <i>E. nitriae</i>	Hippotragini
Gavialidae	cf. Hippotragini gen. et sp. nov.
Gavialidae gen. et sp. nov.	

Miocene localities, the hyaenids are the most numerous in both species and individuals. Three species can be identified: *Hyaenictitherium* cf. *H. hyaenoides*; *Ictitherium* sp., the size of *I. taurinum* but with different tooth morphology; and a third (Hyaenidae indet.) is small, like *Protictitherium*. The felid *Machairodus* cf. *M. giganteus* is represented only by very large mandibles. *Machairodus* is rare in Africa, and until now most of the referred material has been attributed to *Dinofelis*. Mustelids are represented only by a small species of lutrine (otter).

Except for the hominid remains⁵, primates are represented by a single damaged maxilla referred to an indeterminate colobine monkey. Rodent remains are mainly attributable to Murinae, but some Sciuridae (*Xerus* sp.) and Hystricidae (*Hystrix* sp.) are also present. Among aardvarks, cranial and postcranial remains can be referred to *Orycteropus* cf. *O. gaudryi*, a species known in the Late Miocene of the eastern Mediterranean area¹³. A medium-sized equid shows strong affinities with *Hipparion abudhabiense* from the Upper Miocene of the United Arab Emirates¹⁴. Fossil remains of proboscideans are abundant. Numerous teeth of both the gomphotheres *Anancus kenyensis* and the elephantid *Loxodonta* sp. aff. *L. sp. indet.* 'Lukeino stage'¹⁵ have been found.

Suids are represented by *Nyanzachoerus syrticus*, smaller than *N.*

kanamensis but larger than *N. devauxi*. The cheek teeth are very low crowned with robust premolars, including a double-rooted upper premolar P¹. The third molars show primitive features: the talon of upper M³ comprises only one lingually located main cusp but lacks developed accessory cusps; the lower M₃ talonid is also simple. Morphology and size are congruent with *Nyanzachoerus syrticus* from Lothagam (Nawata formation, lower member)¹⁶.

The AU receives its name from the abundance of anthracotheriid fossils (Fig. 2). These remains are referred to the north African *Libycosaurus petrocchii*. Together with the Anthracotheriidae, TM 266 yielded a new species of large hippopotamid, equivalent in size to *Hexaprotodon harvardi* from the Lothagam Nawata formation¹⁷. Nevertheless, mandibular morphology is dominated by a particularly high and massive symphysis, which clearly differs from the *H. harvardi* mandible. This primitive species is absent from the Miocene–Pliocene boundary and Early Pliocene sites in Chad, where a smaller and more-derived species is present.

Two giraffid taxa are represented, a large and a small form. The former, *Sivatherium* cf. *S. hendeyi*, is documented by mandibles and postcranial material. The latter is represented by postcranial bones, tentatively assigned to Giraffidae indet.

The hominid locality has yielded at least five bovid species, three of them being abundant. The smallest is a kob with short horns, similar to *Kobus subdulus* and *K. darti* from Langebaanweg (South Africa)¹⁸, Sahabi (Libya)¹⁹, Manonga (Tanzania)²⁰ and Mpesida (Kenya)²¹, but smaller than in the first two sites. The teeth are more fully reduncine than in Langebaanweg¹⁸ but teeth from this site are perhaps not illustrative of the teeth of primitive reduncines. The other two common taxa are of uncertain tribal affinities. One is attributed to cf. Hippotragini gen. et sp. nov., with a rather primitive skull morphology and long curved horns with a large basal sinus. It is probably an early hippotragine. The teeth are not strongly derived, but would confirm this identification. This species might be the same as ?*Hippotragus* sp. from Sahabi. The last of the three common taxa is a little larger, with massive, almost parallel, slightly curved and not very long horn cores, no sinus in the pedicle and large supraorbital pits. There are some similarities with the ovibovine *Palaeoryx* from the eastern Mediterranean Turolian, but also some differences, which make a close relationship unlikely. For the time being, it is referred to aff. *Palaeoryx* sp. The two less-abundant species are an antilopine and a bovine with primitive teeth. The bovid fauna from the Lothagam Nawata formation (lower member)²² looks more modern than that of TM 266. It is dominated by the modern genus *Aepyceros* and by Boselaphini, which are absent from TM 266 but are known in other African sites close in age to the Miocene–Pliocene boundary, whereas two of the dominant TM 266 bovids (cf. Hippotragini and Ovibovini indet. aff. *Palaeoryx* sp.) are not present at Lothagam.

Biochronology

In the AU of TM 266, *Nyanzachoerus syrticus* is associated with an anthracotheriid (*Libycosaurus*). In younger sites from the Djurab area that have yielded *Nyanzachoerus kanamensis* (such as Kossom Bougoudi)⁴, anthracotheriids are always absent. The evolutionary level of the upper and lower third molars of *Nyanzachoerus syrticus* from the hominid locality is very similar to that of related remains from the Lothagam Nawata formation¹⁶. The age of the Nawata formation fauna is estimated at about 5.2–7.4 Myr (ref. 23). The biochronological age estimate for the hominid locality is also supported by the evolutionary levels of several other mammalian groups (Anthracotheiidae, Proboscidea, Equidae and Bovidae). The *Libycosaurus* anthracotheriids are similar to those from Sahabi²⁴; the fauna from this site is consistent with an Upper Miocene age²⁵. Anthracotheriids are unknown from this time in East Africa and the remains from Libya and Chad may represent their last known occurrences in Africa.

The co-occurrence of *Anancus kenyensis* with *Loxodonta* sp. indet.

'Lukeino stage' is also seen in the Lukeino Formation, Kenya (~6 Myr)^{26,27}. However, the TM 266 proboscideans show more primitive characters than those from Lukeino^{15,28} (thicker enamel and lower laminar frequency for *Loxodonta* sp. aff. *L. sp. indet.* 'Lukeino stage'; a simpler crown pattern with less numerous tubercles and thicker enamel for *A. kenyensis*). These features support an older biochronological age for the TM 266 hominid site. This biochronological estimate is also supported by the presence of the equid *Hipparion* cf. *H. abudhabiense*, a species described from the Lower–Middle Turolian of Abu Dhabi¹⁴. Among carnivores, the co-occurrence of *Machairodus* cf. *M. giganteus*, small *Hyaenictitherium* and *Ichittherium* sp. is congruent with a Late Miocene age, equivalent to the European faunas of the Middle–Upper Turolian. Moreover, the five associated taxa of primitive bovids are also compatible with an Upper Miocene age. Among them, the two larger taxa do not fit easily into modern African bovid tribes, they suggest an earlier age than the Miocene–Pliocene boundary.

This evidence suggests that TM 266 is older than the Lukeino Formation^{26,27}, and may be equivalent in age with the base of the fossiliferous levels of the Nawata formation at Lothagam²³. It is hoped that continuing studies on the fauna will clarify the precise chronological position of TM 266, but present evidence suggests that the faunal remains, including the hominid remains, were deposited between 6 and 7 Myr ago.

Environmental reconstruction

The diversity of aquatic and amphibious forms clearly demonstrates the presence of aquatic environments. Among fish, *Hydrocynus* (tiger fish) is a strict piscivore, hunting by sight in deep and well-oxygenated waters; many of the specimens are longer than 1 m, indicating extensive aquatic habitats. The frequency of piscivorous crocodilians, including *Euthecodon* and a new species of gavial, also clearly indicates large and permanent water bodies. Swampy, well-vegetated areas are suggested by the presence of the fish *Polypterus*, which can tolerate very poorly oxygenated waters. Likewise, *Gymnarchus* lives in swampy and turbid waters where its electrical sensorial system is advantageous. Moreover, there is a great abundance of taxa related to amphibious and bank habitats, particularly the well-preserved anthracotheriids and hippopotamids (including some complete skeletons), otters, trionychid turtles and the snake *Python* cf. *P. sebae*. The liana-like character of the papilionoid plants from TM 266 is compatible with a gallery forest bordering a lake.

The bovids represent about 55% of mammal remains from the TM 266 locality and amphibious mammals represent about 28%. The relatively high-crowned teeth of all the bovids and the lack of tragelaphines and boselaphines are strong evidence of open grassland. Thus, the faunal assemblage from the hominid site is compatible with a diversity of habitats including grassland (bovids), wooded savannah (proboscideans and giraffids), fresh water (fish, turtles, snakes, crocodilians, anthracotheriids, hippos and otters) and probably some gallery forest (primates).

The oldest known East African hominids (*Orororrin*, *Ardipithecus*) are contemporary with faunas associated with wooded environments^{26,29,30}. Younger australopithecines lived in a wider range of habitats³¹. In contrast, the TM 266 vertebrate fauna contemporary of the Toros-Menalla hominid suggests a mosaic of environments from gallery forest at the edge of a lake area to a dominance of large savannah and grassland. Determining the precise habitat of the TM 266 hominid locality among the mosaic of environments available to it constitutes a research challenge to be met by further laboratory and field studies currently in progress. □

Received 13 March; accepted 27 May 2002; doi:10.1038/nature00880.

1. Brunet, M. *et al.* The first australopithecine 2,500 kilometres west of the Rift Valley (Chad). *Nature* **378**, 273–275 (1995).
2. Brunet, M. *et al.* *Australopithecus bahrelghazali*, une nouvelle espèce d'Hominidé ancien de la région de Koro Toro (Tchad). *C.R. Acad. Sci. Paris* **322**, 907–913 (1996).
3. Brunet, M. *et al.* Tchad: découverte d'une faune de mammifères du Pliocène inférieur. *C.R. Acad. Sci.*

Paris **326**, 153–158 (1998).

4. Brunet, M. & MPFT Chad: Discovery of a vertebrate fauna close to the Mio-Pliocene boundary. *J. Vert. Paleontol.* **20**, 205–209 (2000).
5. Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418**, 145–151 (2002).
6. Ghienne, J. F., Schuster, M., Bernard, A., Düringer, P. & Brunet, M. The Holocene giant Lake Chad revealed by Digital Elevation Models. *Quat. Int.* **87**, 81–85 (2002).
7. Hunter, R. E. Basic types of stratification in small eolian dunes. *Sedimentology* **24**, 361–387 (1977).
8. Hunter, R. E. & Kocurek, G. An experimental study of subaqueous slipface deposition. *J. Sediment. Petrol.* **56**, 387–394 (1986).
9. Düringer, P. *et al.* First discovery of fossil dung beetle brood balls and nests in the Chadian Pliocene australopithecine levels. *Lethaia* **33**, 277–284 (2000).
10. Schuster, M. *et al.* Découverte de termitières fossiles dans les sites à vertébrés du Pliocène tchadien: description, identification et implications paléocologiques. *C.R. Acad. Sci. Paris* **331**, 15–20 (2000).
11. Seignobos, C. in *Datation et Chronologie dans le Bassin du lac Tchad* Série Colloques & Séminaires ORSTOM (eds Barreteau, D. & Von Graffenried, C.) 165–182 (ORSTOM, Paris, 1993).
12. Stewart, K. M. in *Neogene Paleontology of the Manonga Valley, Tanzania* Topics in Geobiology Vol. 14 (ed. Harrison, T.) 333–349 (Plenum, New York, 1997).
13. Bonis, de L., Bouvain, G., Geraads, D., Koufos, G. D. & Sen, S. The first aardvarks (Mammalia) from the late Miocene of Macedonia, Greece. *Neues Jahrbuch Geol. Paläont. Abhandlungen* **194**, 343–360 (1994).
14. Eisenmann, V. & Whybrow, P. J. in *Fossil Vertebrates of Arabia* (eds Whybrow, P. J. & Hill, A.) 234–253 (Yale Univ. Press, New Haven, 1999).
15. Tassy, P. in *Geology of the Albertine Rift Valley, Uganda–Zaire* Vol. II (eds Senut, B. & Pickford, M.) 217–257 (CIFEG Occas. Publ., Orléans, 1994).
16. Harris, J. M. & White, T. D. Evolution of the Plio-Pleistocene African Suidae. *Trans. Am. Phil. Soc.* **69**(2), 1–128 (1979).
17. Coryndon, S. C. The taxonomy and nomenclature of the Hippopotamidae (Mammalia, Artiodactyla) and a description of two new fossil species. *Proc. Koninklijke Neder. Akad. Wetenschappen B* **80**, 61–88 (1977).
18. Gentry, A. W. Fossil Bovidae (Mammalia) from Langebaanweg, South Africa. *Ann. S. Afr. Mus.* **79**(8), 213–337 (1980).
19. Lehmann, U. & Thomas, H. in *Neogene Paleontology and Geology of Sahabi* (eds Boaz, N. T., El-Arnauti, A., Gaziry, A. W., de Heinzelin, J. & Boaz, D. D.) 323–335 (Liss, New York, 1987).
20. Gentry, A. W. in *Neogene Paleontology of the Manonga Valley, Tanzania* Topics in Geobiology Vol. 14 (ed. Harrison, T.) 107–135 (Plenum, New York, 1997).
21. Thomas, H. in *Proc. 8th Pan African Congress of Prehistory* (eds Leakey, L. E. F. & Ogot, B. A.) 82–91 (Nairobi, 1977).
22. Leakey, G. M. *et al.* Lothagam: a record of faunal change in the Late Miocene of East Africa. *J. Vert. Paleontol.* **16**, 556–570 (1996).
23. McDougall, I. & Feibel, C. S. Numerical age control for the Miocene–Pliocene succession at Lothagam, a hominoid-bearing sequence in the northern Kenya Rift. *J. Geol. Soc. Lond.* **156**, 731–745 (1999).
24. Gaziry, A. W. in *Neogene Paleontology and Geology of Sahabi* (eds Boaz, N. T., El-Arnauti, A., Gaziry, A. W., de Heinzelin, J. & Boaz, D. D.) 287–302 (Liss, New York, 1987).
25. Geraads, D. Vertébrés fossiles du Miocène supérieur du Djebel Krechem el Artsouma (Tunisie centrale). Comparaisons biostratigraphiques. *Géobios (Lyon)* **22**, 777–901 (1989).
26. Pickford, M. & Senut, B. The geological and faunal context of Late Miocene hominid remains from Lukeino, Kenya. *C.R. Acad. Sci. Paris* **332**, 145–152 (2001).
27. Deino, A. L., Tauxe, L., Monaghan, M. & Hill, A. ⁴⁰Ar/³⁹Ar geochronology and paleomagnetic stratigraphy of the Lukeino and lower Chemeron Formations at Tabarin and Kapcheberek, Tugen Hills, Kenya. *J. Hum. Evol.* **42**, 117–140 (2002).
28. Tassy, P. Nouveaux Elephantoidea (Mammalia) dans le Miocène du Kenya Cahiers de Paléontologie, Travaux Paléontologie Est-africaine (CNRS, Paris, 1986).
29. WoldeGabriel, G. *et al.* Ecological and temporal placement of early Pliocene hominids at Aramis, Ethiopia. *Nature* **371**, 330–333 (1994).
30. WoldeGabriel, G. *et al.* Geology and palaeontology of the Late Miocene Middle Awash valley, Afar rift, Ethiopia. *Nature* **412**, 175–178 (2001).
31. Leakey, M. D. & Harris, J. M. (eds) *Laetoli: A Pliocene Site in Northern Tanzania* 524–531 (Clarendon, Oxford, 1987).

Acknowledgements

We thank the Chadian authorities (Ministère de l'Éducation Nationale de l'Enseignement Supérieur et de la Recherche, the Université de N'djamena, and CNAR). We extend gratitude for their support to the French Ministries, Ministère de l'Éducation Nationale (Faculté des Sciences, Université de Poitiers), Ministère de la Recherche (CNRS) and Ministère des Affaires Étrangères (Direction de la Coopération Scientifique, Universitaire et de Recherche, Paris, and SCAC Ambassade de France à N'djamena), to the Région Poitou-Charentes, le Département de la Vienne, the Association pour le Prix scientifique Philip Morris, and also to the Armée Française (MAM and Epervier) for logistic support. Many thanks to the National Museum of Kenya, National Museum of Ethiopia and University of California, who allowed us access to their collections. We especially thank A. Djimdoumalbaye, A. Beauvilain, J.-C. Bertrand, F. Gongdibé, M. Adoum and X. Valentin for field and technical work and all the other MPFT participants to our field missions. We thank F. Clark Howell, Y. Haile-Selassie, M. G. Leakey, D. Lieberman, M. Muungu, D. Pilbeam and T. White for insights and comments. We are most grateful to G. Florent for administrative guidance of the MPFT.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.V. (e-mail: patrick.vignaud@univ-poitiers.fr).

Possible *in situ* formation of meteoritic nanodiamonds in the early Solar System

Z. R. Dai*, J. P. Bradley*†, D. J. Joswiak‡, D. E. Brownlee‡, H. G. M. Hill§ & M. J. Genge||

* School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

† Department of Astronomy, University of Washington, Seattle, Washington 98195, USA

‡ Laboratory for Extraterrestrial Physics, Code 691, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

§ Department of Mineralogy, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

Grains of dust that pre-date the Sun provide insights into their formation around other stars and into the early evolution of the Solar System^{1–4}. Nanodiamonds recovered from meteorites, which originate in asteroids, have been thought to be the most abundant type of presolar grain^{3,4}. If that is true, then nanodiamonds should be at least as abundant in comets, because they are thought to have formed further out in the early Solar System than the asteroid parent bodies, and because they should be more pristine^{5–7}. Here we report that nanodiamonds are absent or very depleted in fragile, carbon-rich interplanetary dust particles, some of which enter the atmosphere at speeds within the range of cometary meteors^{8,9}. One interpretation of the results is that some (perhaps most) nanodiamonds formed within the inner Solar System and are not presolar at all, consistent with the recent detection of nanodiamonds within the accretion discs of other young stars¹⁰. An alternative explanation is that all meteoritic nanodiamonds are indeed presolar, but that their abundance decreases with heliocentric distance, in which case our understanding of large-scale transport and circulation within the early Solar System is incomplete¹¹.

The normalized average abundance of nanodiamonds in carbonaceous chondrites of types CI and CM (from the inner main belt of the asteroids) is 750–1,500 p.p.m. (refs 1, 2, 7). Most meteoritic nanodiamonds are between 1 and 10 nm in diameter, with a lognormal size distribution and an average diameter of ~3 nm (refs 2, 12). Bulk extracts of (colloidal) nanodiamonds are isolated from meteorites by a multi-step chemical (acid digestion) process that has been described as “burning down the haystack to find the needle”⁴. Individual nanodiamonds can be directly observed within electron-transparent sections of meteorites, micrometeorites, and interplanetary dust particles (IDPs; from comets and asteroids) mounted on gold-mesh transmission electron microscope (TEM) grids following *in situ* acid etching¹³. We have examined sections of IDPs, meteorites and micrometeorites (see Table 1) using TEM.

Figure 1 shows bright-field images of a thin section of IDP W7027A-8D before and after *in situ* acid etching. Only disordered carbon was observed in the etched IDP sections, but the meteorite and micrometeorite sections also contain a graphitic carbon component. Figure 2 shows experimental lattice-fringe images of nanodiamonds in thin sections of a meteorite, a micrometeorite and four IDPs. Figure 3 shows simulated lattice-fringe images of diamond, gold (Au), palladium (Pd) and tetragonal palladium fluoride (t-PdF₂).

Although bulk samples of meteoritic nanodiamonds can be readily identified in the TEM^{2,12}, we found that individual nanodiamonds (typically with <3,000 carbon atoms) are more difficult to identify because they are supported on (much thicker) carbon

Table 1 Samples examined in this work

Samples without nanodiamonds*	Samples with nanodiamonds
W7207A8D, U217B19, U2073B4C, U230C4D, U216B4	Murchison†, Orgueil (2 samples)‡, AMMSRU2-2§, AMM94-2/14§, RB12A44-3 ¶, U230C1GB¶, U220GCA (3 samples) ¶, W7110A2ED¶

One of the three nanodiamond-containing cluster IDPs (RB12A44-3) has CM mineralogy, and is therefore an asteroidal particle¹⁵. (In an earlier study, the mineralogy of another giant cluster particle (L2008#5) was linked to an asteroidal origin¹⁶.) Because Murchison (CM) and Orgueil (CI) meteorites, the two CM micrometeorites, and the CM IDP RB12A44-3 are compact objects with layer lattice silicates consistent with an asteroidal origin, it is expected that they contain nanodiamonds. The other three cluster particles (U2-20GCA, U2-30C-1G-B and W7110A-2E-D) are fragile objects that contain anhydrous silicates consistent with a cometary origin. However, it is equally plausible that such anhydrous cluster particles are captured from low-speed orbits because they are from ‘comet-like’ asteroids or Kuiper-belt objects that are not sampled by meteorites²⁵.

* All five are non-cluster IDPs.

† CM (type 2) carbonaceous chondrite.

‡ CI (type 1) carbonaceous chondrite.

§ ‘CM-like’ polar micrometeorite.

|| ‘CM-like’ IDP.

¶ Fragment of a cluster particle.

substrates and, in several etched sections, it was necessary to distinguish them from nanoparticle trace contaminants like gold and palladium fluoride. The gold is derived from partial dissolution and precipitation of gold from the (gold) TEM grid during the *in situ* HF acid etching procedure, and palladium metal was applied to some particles as a conductive coating before thin sectioning. Because gold and palladium have lattice-fringe spacings and lattice structures similar to those of nanodiamonds (Fig. 3), precise *in situ* image calibration (using gold and graphitic carbon) was required.

We identified nanodiamonds within etched thin sections of the Murchison (CM) and Orgueil (CI) meteorites, two ‘CM-like’ micrometeorites, and four IDPs. Nanodiamonds were not identified in any of the smaller non-cluster IDPs (Table 1). All of the nanodiamond-containing IDPs are pieces of so-called ‘giant cluster particles’, unusually large (50–500 µm diameter) IDPs that break

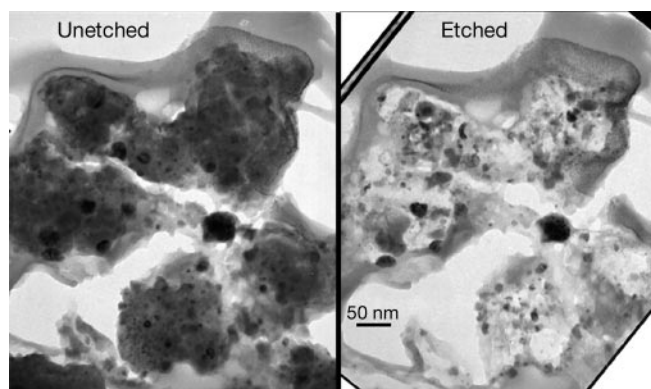


Figure 1 Bright-field images of a thin section of chondritic IDP W7027A8D before and after *in situ* acid etching. The unetched section (left) contains silicate grains (enstatite, forsterite, glass with embedded metal and sulphides (GEMS)²⁶, FeNi metal, FeNi sulphides, and carbonaceous material. The etched section (right) contains only carbonaceous material (grey matrix) and FeNi sulphides (dark inclusions). IDPs collected in the stratosphere are from comets and asteroids^{8,9,15}. Asteroidal IDPs so far identified are similar to (asteroidal) meteorites in that they are compact objects that contain hydrated layer lattice silicates¹⁵. Cometary IDPs are porous fragile objects that contain anhydrous silicates^{8,9}. Chondritic IDPs contain 2–5 times more carbon than do chondritic meteorites, and some of them are the most chemically and isotopically primitive meteoritic materials available for laboratory investigation^{14,27,28}. We also etched sections of fine-grained matrix from the Murchison (CM) and Orgueil (CI) carbonaceous chondrites, two polar CM-like (asteroidal) micrometeorites, and nine chondritic IDPs (Table 1). The mineralogy and petrography of sections before and after etching were examined using 200-kV (HF2000 and JEOL 2010) and 400-kV (JEOL 4000EX) TEMs. The compositions of the sections were measured using X-ray energy-dispersive spectroscopy and electron energy-loss spectroscopy, and the structures of individual nanocrystals among the etched residues were investigated using lattice-fringe imaging.

† Present address: Institute for Geophysics & Planetary Physics, Lawrence Livermore National Laboratory, Mail Stop L-413, 7000 East Avenue, Livermore, California 94550, USA.

into multiple fragments during collection. The relationship between large, nanodiamond-containing cluster IDPs and smaller, non-cluster IDPs is unclear. It has been suggested that cluster particles are more isotopically primitive than smaller non-cluster IDPs¹⁴, although as both classes of IDP have similar compositions, mineralogy and petrography, both could be derived from either comets and/or asteroids. In general, cluster particles appear less strongly heated than is predicted by models of atmospheric-entry heating, implying that they are captured from low-speed asteroidal orbits^{15,16}. The spectral characteristics of cometary nuclei suggest that they may be related to some asteroidal objects now found within the inner Solar System¹⁷. In other words, the difference between some comets and some primitive asteroids may simply be that of orbital parameter.

Although nanodiamond-rich meteoritic acid residues are not pure (they also contain amorphous carbon, organic carbon, oxides, and dissolved silicate residues), and their bulk (¹³C/¹²C) isotopic composition is within the range of Solar System materials¹⁸, gases (for example, N and Xe) within the acid residues have non-solar isotopic compositions¹⁹. During stepwise combustion, these gases are released at the combustion temperature of diamond, indicating that a fraction of the nanodiamonds are almost certainly presolar stardust grains. But in contrast to other presolar grains (for example, SiC), the isotopic compositions of individual nanodiamonds have never been measured (because they are too small). Individual nanodiamonds probably exhibit a range of ¹³C/¹²C compositions, and, although a non-solar value unambiguously establishes presolar origin, the reverse is not necessarily true—that is, presolar nanodiamonds may exhibit both solar and non-solar isotopic compositions. This difficulty in identifying presolar components highlights the importance of coupling laboratory measurements with astronomical observations. The identification of

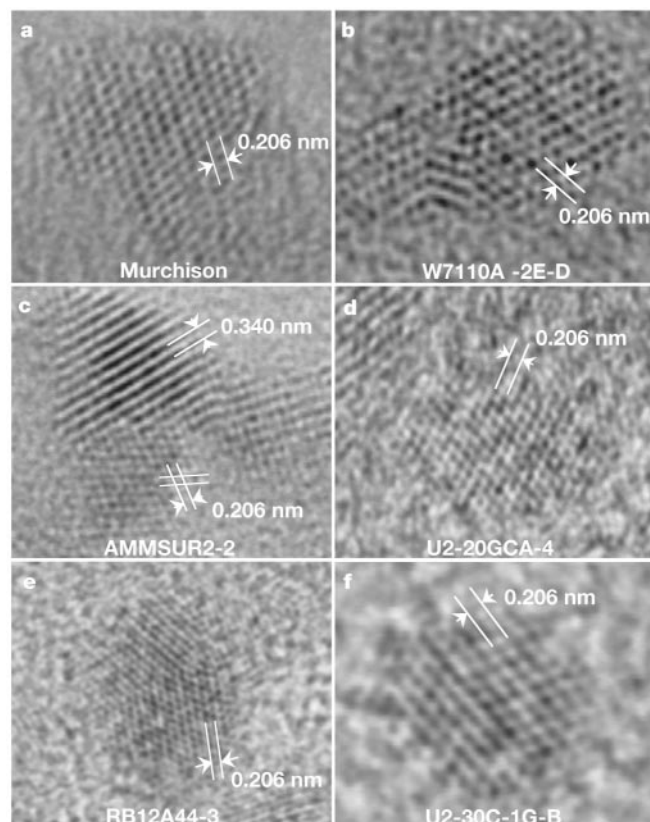


Figure 2 Experimental lattice-fringe images of nanodiamonds. **a**, Murchison (CM) meteorite; **b**, IDP W7110A2ED; **c**, micrometeorite AMMSRU2-2; **d**, IDP U220GCA; **e**, RB12A443; and **f**, U230C1GD.

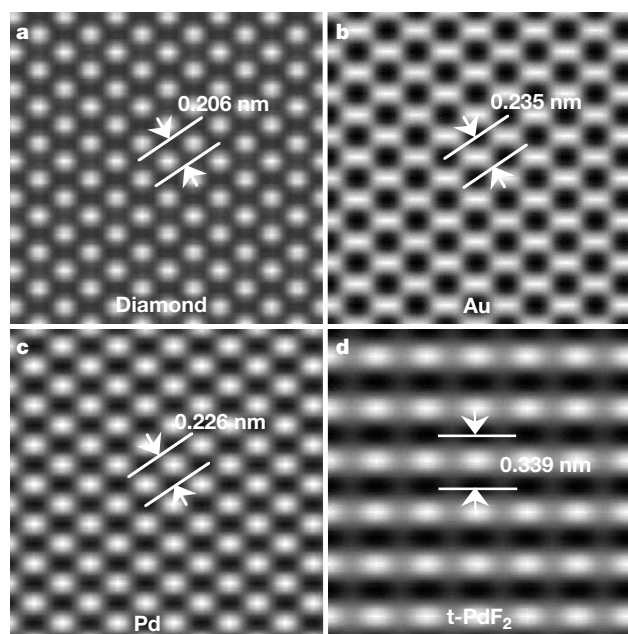


Figure 3 Simulated lattice-fringe images. **a**, Diamond; **b**, gold; **c**, palladium; and **d**, tetragonal palladium fluoride. In this orientation the Au, Pd and t-PdF₂ {111} spacings are 0.235 nm, 0.226 nm and 0.339 nm respectively, and they differ from the 0.206-Å nanodiamond {111} spacing by $\geq 9\%$. Orientation is [110] in all images.

presolar SiC grains in meteorites is secure, not only because of their non-solar ¹³C/¹²C compositions, but also because astronomical spectral features due to SiC grains are observed in the outflows of evolved carbon-rich stars²⁰. In contrast, unambiguous evidence of nanodiamonds in the interstellar medium has yet to be rigorously established^{10,21,22}.

We have identified a class of meteoritic materials in which nanodiamonds are either depleted or absent. Assuming that this trend continues as more IDPs are examined, and regardless of whether these particles are from comets or 'comet-like' outer asteroids, the discovery of a class of cosmically primitive, carbon-rich IDPs deficient in nanodiamonds is unexpected. The simplest explanation is that at least some meteoritic nanodiamonds formed in the inner solar nebula, and not in a presolar environment. But there is evidence that meteoritic nanodiamonds formed under reducing conditions², and conditions in the solar nebula are believed to have been oxidizing²³, although nanodiamonds can also form under oxidizing conditions²⁴. An alternative explanation is that nanodiamonds are indeed presolar, but that their abundance decreases with heliocentric distance. In this model, much of the solid matter of the outer Solar System is processed inner Solar System material that was circulated through the accretion disk and then transported outwards¹¹. Perhaps nanodiamonds are not uniformly distributed among carbon-rich non-cluster IDPs, or their (cometary or asteroidal) parent bodies are not as primitive as we have assumed. Any one of these explanations has profound implications concerning our understanding of the early Solar System. □

Received 29 January; accepted 30 May 2002; doi:10.1038/nature00897.

- Hoppe, P. & Zinner, E. Presolar dust grains from meteorites and their stellar sources. *J. Geophys. Res.* **510**, 10371–10385 (2000).
- Daulton, T. L., Eisenhour, D. D., Bernatowicz, T. J., Lewis, R. S. & Buseck, P. R. Genesis of presolar diamonds: comparative high-resolution transmission electron microscopy study of meteoritic and terrestrial nano-diamonds. *Geochim. Cosmochim. Acta* **60**, 4853–4872 (1996).
- Anders, E. & Zinner, E. Interstellar grains in primitive meteorites: diamond, silicon carbide, and graphite. *Meteoritics* **28**, 490–514 (1993).
- Bernatowicz, T. J. & Walker, R. M. Ancient stardust in the laboratory. *Phys. Today* **50**, 26–32 (1997).
- Farinella, P., Gonzi, R. & Froeschle, C. The injection of asteroid fragments into resonances. *Icarus* **101**, 174–187 (1993).
- Greenberg, J. M. in *Comets* (ed. Wilkening, L. L.) 131–162 (Univ. Arizona Press, Tucson, 1982).

7. Huss, G. in *Astrophysical Implications of the Laboratory Study of Presolar Materials* (eds Bernatowicz, T. & Zinner, E.) 721–748 (AIP Press, New York, 1997).
8. Bradley, J. P. & Brownlee, D. E. Cometary particles: thin-sectioning and electron-beam analysis. *Science* **231**, 1542–1544 (1984).
9. Brownlee, D. E. *et al.* Identification of individual cometary IDPs by thermally stepped He release. *Lunar Planet Sci.* **XXVI**, 183–184 (1995).
10. Van Kerckhoven, C., Tielens, A. G. G. M. & Waelkens, C. Nanodiamonds around HD 97048 and Elias 1. *Astron. Astrophys.* **384**, 568–584 (2002).
11. Hill, H. G. M., Grady, C. A., Nuth, J. A., Hallenbeck, S. L. & Sitko, M. L. Constraints on nebular dynamics and chemistry based on observations of annealed magnesium silicate grains in comets and in disks around Herbig Ae and Be stars. *Proc. Natl Acad. Sci. USA* **98**, 2182–2187 (2001).
12. Bernatowicz, T. J. & Gibbons, P. C. Electron energy-loss spectroscopy of interstellar diamonds. *Astrophys. J.* **359**, 246–255 (1990).
13. Brownlee, D. E., Joswiak, D. J., Bradley, J. P., Gezo, J. C. & Hill, H. G. M. Spatially resolved acid dissolution of IDPs: the state of carbon and the abundance of diamonds in the dust. *Lunar Planet. Sci.* **XXXI**, 1921–1922 (2000).
14. Messenger, S. Identification of molecular cloud material in interplanetary dust particles. *Nature* **404**, 968–971 (2000).
15. Bradley, J. P. & Brownlee, D. E. An interplanetary dust particle linked to type CM meteorites and an asteroidal origin. *Science* **251**, 489–496 (1992).
16. Thomas, K. L. *et al.* An asteroidal breccia: The anatomy of a cluster IDP. *Geochim. Cosmochim. Acta* **59**, 2797–2815 (1995).
17. Luu, J. X. Spectral diversity among the nuclei of comets. *Icarus* **104**, 138–148 (1993).
18. Russell, S. S., Arden, J. W. & Pillinger, C. T. A carbon and nitrogen isotope study of diamond from primitive meteorites. *Meteorit. Planet. Sci.* **31**, 343–355 (1996).
19. Huss, G. R. & Lewis, R. S. Noble gases in presolar diamonds. I. Three distinct components and their implications for diamond origins. *Meteoritics* **29**, 791–810 (1994).
20. Sylvester, R. J. in *Solid Interstellar Matter: The ISO Revolution* (eds d'Hendecourt, L., Joblin, C. & Jones, A.) 263–276 (Springer, New York, 1999).
21. Hill, H. G. M., Jones, A. P. & d'Hendecourt, L. B. Diamonds in carbon-rich proto-planetary nebulae. *Astron. Astrophys.* **336**, L41–L44 (1998).
22. Guillois, O., Ledoux, G. & Reynaud, C. Diamond infrared emission bands in circumstellar media. *Astrophys. J.* **521**, L133–L136 (1999).
23. Lodders, K. & Palme, H. in *Protostars and Planets IV* (eds Krot, A. N. & Fegley, B.) 1019–1053 (Univ. Arizona Press, Tucson, 2000).
24. Howard, W. *et al.* Synthesis of diamond powder in acetylene oxygen plasma. *J. Appl. Phys.* **68**, 1247–1251 (1990).
25. Liou, J.-C., Zook, H. A. & Dermott, S. F. Kuiper belt dust grains as a source of interplanetary dust particles. *Icarus* **124**, 429–440 (1996).
26. Bradley, J. P. Chemically anomalous, pre-accretionally irradiated grains in interplanetary dust particles from comets. *Science* **265**, 925–927 (1992).
27. Thomas, K. L., Blanford, G. E., Keller, L. P., Klöck, W. & McKay, D. S. Carbon abundance and silicate mineralogy of anhydrous interplanetary dust particles. *Geochim. Cosmochim. Acta* **57**, 1551–1566 (1993).
28. Schramm, L. S., Brownlee, D. E. & Wheelock, M. M. Major element compositions of stratospheric micrometeorites. *Meteoritics* **24**, 99–112 (1989).

Acknowledgements

We thank T. Bernatowicz for bulk nanodiamond extracts, A.W. Phelps for discussions, and G. Huss for comments and suggestions. This work was supported by NASA and the Georgia Tech Electron Microscopy Center. H.G.M.H. acknowledges support from the NAS/NRC RRA programme.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.P.B. (e-mail: jbradley@igpp.uclnl.org).

Phonon-enhanced light–matter interaction at the nanometre scale

R. Hillenbrand, T. Taubner & F. Keilmann

Max-Planck-Institut für Biochemie, Abteilung Molekulare Strukturbiologie, 82152 Martinsried & Center for NanoScience, Ludwig-Maximilians-Universität, 80799 München, Germany

Optical near fields exist close to any illuminated object. They account for interesting effects such as enhanced pinhole transmission¹ or enhanced Raman scattering enabling single-molecule spectroscopy². Also, they enable high-resolution (below 10 nm) optical microscopy^{3–6}. The plasmon-enhanced near-field coupling

between metallic nanostructures^{7–9} opens new ways of designing optical properties^{10–12} and of controlling light on the nanometre scale^{13,14}. Here we study the strong enhancement of optical near-field coupling in the infrared by lattice vibrations (phonons) of polar dielectrics. We combine infrared spectroscopy with a near-field microscope that provides a confined field to probe the local interaction with a SiC sample. The phonon resonance occurs at 920 cm^{−1}. Within 20 cm^{−1} of the resonance, the near-field signal increases 200-fold; on resonance, the signal exceeds by 20 times the value obtained with a gold sample. We find that phonon-enhanced near-field coupling is extremely sensitive to chemical and structural composition of polar samples, permitting nanometre-scale analysis of semiconductors and minerals. The excellent physical and chemical stability of SiC in particular may allow the design of nanometre-scale optical circuits for high-temperature and high-power operation.

A recent approach to confining and guiding light uses the strong near fields of closely spaced metallic nanostructures^{13,14}. In these structures, a collective motion of conduction electrons (the surface plasmon polariton¹⁵, referred to here as “plasmon”) becomes resonantly excited by visible light. The interparticle coupling by plasmons provides field confinement and photon transport over mesoscopic distances. An infrared counterpart to the surface plasmon polariton is the surface phonon polariton (referred to here as “phonon”), which has weaker damping¹⁵ and thus offers the advantage of stronger and sharper optical resonances. We demonstrate this for the small-particle (Fröhlich¹⁶) resonance, by calculating the enhanced optical field E_{loc} at the surface of an illuminated sphere with radius a and dielectric function ϵ_p (the incident field is E_{in}). In the electrostatic approximation (which is valid as we assume a particle size much smaller than the wavelength) the sphere's polarizability is given by

$$\alpha = 4\pi a^3 (\epsilon_p - \epsilon_m) / (\epsilon_p + 2\epsilon_m) \quad (1)$$

where ϵ_m is the dielectric function of the embedding medium¹⁷. The local field enhancement $E_{\text{loc}}/E_{\text{in}} \propto \alpha$ has a resonance at a frequency determined by $\text{Re}[\epsilon_p(\omega)] = -2\epsilon_m$ (note this condition can be fulfilled likewise by phonons, plasmons or excitons). By inserting dielectric data of metals¹⁸ and SiC¹⁹ into equation (1), we obtain the curves shown in Fig. 1, which clearly demonstrate that the infrared phonon resonance of a small polar-dielectric particle is much stronger and sharper than the plasmon resonance of an equally small metal particle. Accordingly, we propose phonon photonics as a concept in nanotechnology analogous to plasmon photonics. Compared to plasmon resonance (applied, for example, in small-particle sensors²⁰), the sharper phonon resonance enables an even higher sensitivity to environmental changes affecting ϵ_m , such as molecular adsorption. Phonon resonance also responds to changes in the lattice parameters affecting ϵ_p . This could be used to detect variations in local temperature and pressure. The necessary infrared frequencies could be supplied by quantum cascade lasers^{21,22}, which

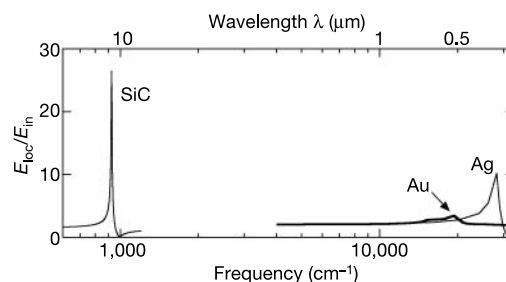


Figure 1 Calculated field enhancement $E_{\text{loc}}/E_{\text{in}}$ due to Fröhlich resonance at the surface of a 10-nm-diameter sphere. The infrared lattice vibration (phonon) of a polar dielectric (SiC) induces a considerably stronger near-field amplitude than the green/blue electronic excitation (plasmon) of a noble metal (Au, Ag).

are readily tailored to match the characteristic phonon frequencies of polar materials that include III–V and II–VI semiconductors.

Here we focus on the phonon enhancement of the optical interaction between two objects at a nanometric distance from each other. Localized optical interaction has become accessible to experimentation by the recent advance of the ‘apertureless’, scattering-type scanning near-field optical microscope (s-SNOM)^{3–6,23,24}, in which near-field coupling is the essential contrast mechanism^{5,6,23,24}. We therefore consider the configuration^{15,25} of a small sphere with radius a at a small distance z from a plane sample (Fig. 2a inset), the former representing the probing tip of the s-SNOM. The electric field is polarized perpendicularly to the sample surface. The sphere induces an image dipole in the sample, with polarizability $\alpha\beta$ where

$$\beta = (\epsilon_s - 1)/(\epsilon_s + 1) \quad (2)$$

is the surface response function, ϵ_s being the dielectric function of the sample. Note that the image dipole can become resonant by the dielectric value of the sample alone, at $\text{Re}[\epsilon_s(\omega)] = -1$, independently of whether the sphere is resonant or not. The near-field interaction between both dipoles is described by an effective polarizability, in electrostatic approximation⁶:

$$\alpha_{\text{eff}} = \frac{\alpha(1 + \beta)}{1 - \frac{\alpha\beta}{16\pi(a+z)^3}} \quad (3)$$

A resonant near-field interaction thus originates from the dielectric resonance of either a polar sphere (equation (1)) or a polar sample (equation (2)). The interaction plays no key role in the generation of the resonance, but modifies strength and spectral position. Without losing generality, we can show the phonon-enhanced interaction by considering the case of a resonant sample material. This choice simplifies an experimental comparison between a phonon-active material (SiC) and a non-resonant reference material (Au). The near-field interaction (equation (3)) is experimentally accessible by measuring the scattered intensity S of the coupled system, which depends on α_{eff} as $S \propto |\alpha_{\text{eff}}|^2$. An evaluation of S for Au and SiC samples is given in Fig. 2a, with Pt as (non-resonant) sphere material. The result for SiC exhibits a very sharp resonance at 930 cm^{-1} (where $\text{Re}[\epsilon_s(\omega)] \approx -1.7$), exceeding the value of Au by nearly two orders of magnitude, in contrast to the far-field interaction (Fig. 2b). In Fig. 2b, the phonon-induced ‘reststrahlen’ reflectivity band of SiC extends from 790 to 950 cm^{-1} , a region where $\text{Re}(\epsilon_s) < 0$ forbids bulk propagation. The near-field probing singles out a sharp resonance within the phonon band, resulting in a spectral response very different from the far-field response. Such a spectral narrowing had been predicted

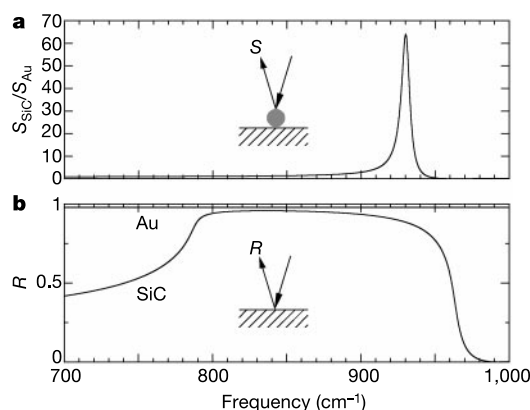


Figure 2 Enhancement and narrowing of a phonon-induced spectral response by the near-field probing process. **a**, Intensity S of scattering from a small Pt probe particle of radius $a \ll \lambda$ in contact with a SiC surface (relative to an Au surface); **b**, far-field reflectivity R of SiC (reststrahlen band) and of Au (nearly perfect mirror).

earlier for the near field of SiC^{25,26}. In the case of ref. 26, it occurs at the asymptotic frequency of the phonon dispersion, given by $\text{Re}[\epsilon_s(\omega)] = -1$, owing to thermally excited phonons.

The experimental proof of the predicted phonon-resonant near-field interaction is carried out by inspecting a flat SiC surface with our infrared s-SNOM^{6,24} (Fig. 3a). Here a sharp tip at the end of a cantilever takes the role of the sphere of the model above. The instrument is basically an atomic force microscope, where the Pt-coated tip (radius $\sim 20 \text{ nm}$) maps the topography of a scanned sample. The tip is illuminated by focused infrared light from a line-tunable CO₂ laser or a quantum cascade laser. E_{in} is polarized in the y - z plane. Back-scattered light is collected through the focusing cassegrainian objective, and detected interferometrically (details to be published elsewhere) to register the scattered amplitude $s = \sqrt{S} \propto |\alpha_{\text{eff}}|$ that directly measures the local near-field interaction between tip and sample surface. To assess the pure near-field interaction, it is essential to suppress unavoidable background light scattered from the tip's shaft and cantilever. Our solution is to use the tapping mode, where the cantilever vibrates at its resonance frequency Ω so that the tip-sample separation z oscillates between 0 and $\sim 40 \text{ nm}$ —that is, with amplitude $\Delta z \approx 20 \text{ nm}$. Owing to the nonlinear z -dependence, the near-field interaction (equation (3)) and thus s become modulated at higher harmonics $n\Omega$ ($n > 1$), but the background scattering is not modulated in this way. Lock-in detection at $n\Omega$ ($n = 2$ used in this work) can therefore suppress the background^{5,23}. The sample is a polished, Acheson grade 6H-SiC crystal partly coated with a 10-nm -thick, nanostructured gold film²⁷. A large number of infrared images are taken repeatedly of the same sample area, which contains both Au and SiC surfaces, at many wavelength settings of the laser.

The infrared image brightness s (Fig. 3c) maps the local near-field coupling between tip and sample, and thus allows us to directly compare different sample materials. Indeed, near the frequency of the predicted phonon resonance (Fig. 2a and image at 929 cm^{-1}) SiC appears much brighter than Au. When changing frequency, the signal changes on SiC but not on Au. A 5% relative frequency shift (image at 978 cm^{-1}) is sufficient to reverse contrast and even lose the signal in the instrument's noise. These two images already demonstrate a strong phonon effect in SiC. To quantify the phonon

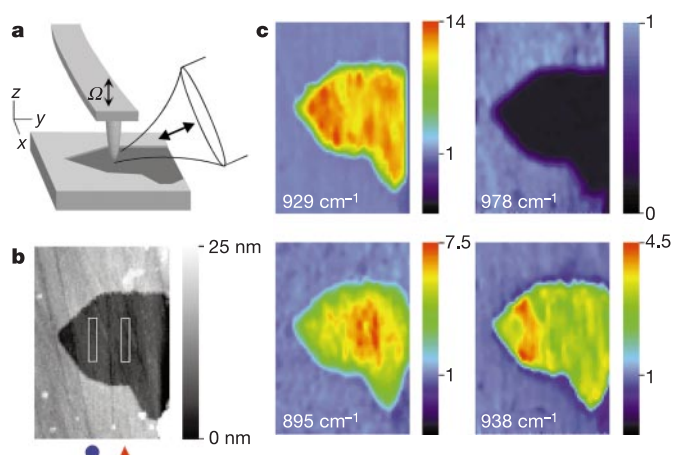


Figure 3 Experimental scheme (**a**) and images (**b,c**) taken with a scattering-type mid-infrared scanning near-field microscope (s-SNOM). The topography (**b**) shows a partly Au-covered SiC sample (image size $1.6 \times 2.3 \mu\text{m}^2$). Two rectangles mark areas used for data extraction. The infrared near-field images (**c**) display the scattering amplitude s taken at different frequencies of illumination as indicated, and are scaled such that Au appears blue. On phonon resonance, the SiC area appears much brighter than the surrounding Au (929 cm^{-1}). The contrast reverses at 978 cm^{-1} . Two examples taken on either side of the resonance (895 cm^{-1} and 938 cm^{-1}) indicate a systematic local variation in the infrared images (see text).

enhancement in detail, we extract two values of s_{SiC} from an infrared image by averaging over rectangular areas indicated in Fig. 3b. We normalize to the value of Au to obtain two values of $s_{\text{SiC}}/s_{\text{Au}}$ from each infrared image, and plot them in a spectral diagram (Fig. 4). The experimental data trace out and resolve a resonance curve over a high dynamic range, which extends at its maximum 20-fold above the near-field coupling observed on Au. The resonance is accompanied by a minimum more than 10 times below the value for Au. Gaps in the spectrum are due to the tuning capability of our gas lasers operating with $^{12}\text{C}^{16}\text{O}_2$ and $^{13}\text{C}^{16}\text{O}_2$ fillings, or to undetectably small signals on SiC as at 968 and 978 cm^{-1} . A near-field image that we have taken with a quantum cascade laser (supplied by C. Gmachl) yields a corresponding value of $s_{\text{SiC}}/s_{\text{Au}} = 0.3$ at 1,360 cm^{-1} .

To compare our findings with theory, we recall the prediction (Fig. 2a) calculated from equation (3), and plot it as the dashed curve in Fig. 4. However, we have to take account of the experimental procedure of tip vibration and second-harmonic signal demodulation. For this simulation, one parameter is to be taken from the experiment—the relative tapping amplitude $\Delta z/a$. Assuming $\Delta z/a = 1$, we calculate the solid curve in Fig. 4, which follows the overall behaviour of the dashed curve but which exhibits a higher dynamic range (a systematic instrumental effect noted before²³) and a broadening. Comparing now the solid curve with the experimental data, we find a strong resonance in both. Discrepancies such as a factor of two in the resonance enhancement and a frequency shift of about 15 cm^{-1} are relatively small. These may call for a refinement of the theory by using a more realistic tip geometry (an example of tip-shape-dependent scattering is found in ref. 28) and a multipolar, electrodynamic formulation. However, the experimental data agree well with the narrow shape and the high dynamic range (three orders of magnitude in amplitude, that is, six orders in intensity) predicted by our model calculation, and thus verify the phonon enhancement of near-field coupling.

The images (Fig. 3c) are direct evidence that infrared spectroscopy is possible with excellent spatial resolution of $<\lambda/100$. This proves that the phonon-enhanced near-field interaction is highly localized. A close inspection of the infrared images reveals a substructure on SiC that depends on frequency (not observed on

Au). The signal at 895 cm^{-1} is about 20% smaller in the left-hand third compared to the rest of the SiC area, but about 20% larger at 938 cm^{-1} . Indeed, the spectra (Fig. 4) extracted from the rectangles in Fig. 3c reveal a systematic frequency-dependent behaviour. At frequencies below resonance, the red triangles are above the blue dots, and this reverses above resonance, indicating a frequency shift of about 2 cm^{-1} between the two sample locations. We take this as experimental evidence of a local variation of the dielectric function, expressing a variation in the physical structure or chemical composition of the sample. This could be due to polishing damage, impurity or doping. Certainly the sharpness of the phonon-enhanced near-field resonance offers extremely high sensitivity in detecting local material inhomogeneities by s-SNOM, an application that reaches far beyond the s-SNOM's demonstrated nanoscale mapping of chemical composition^{5,6,23,24}. The resonance could be used to map crystallinity, dislocations, and profiles of stress or doping. With an experimental steepness of 1.2 dB per cm^{-1} , a single-frequency mapping at 940 cm^{-1} (Fig. 4) would detect, under the assumption of 3% instrumental noise, a local variation of the resonance frequency as small as 0.1 cm^{-1} —that is, a relative change as small as 1×10^{-4} .

A technological application of the present work might be the use of SiC to provide durable, long-term, high-density optical read-only memory (ROM). Phonon enhancement could supply strong signals, enabling a fast near-field read-out of even 10-nm bits, equivalent to a storage density near 1 Tbit inch^{-2} . Also, sharp tips made out of SiC could be durable probes for infrared s-SNOM, with the added benefit that stronger signals might push the resolution limit well below 10 nm. The extreme infrared field strengths to be expected from phonon enhancement in narrow gaps between polar nanoparticles, exceeding those demonstrated for plasmon particles^{2,9}, may bring new opportunities to high-field nonlinear physics. □

Received 27 March; accepted 31 May 2002; doi:10.1038/nature00899.

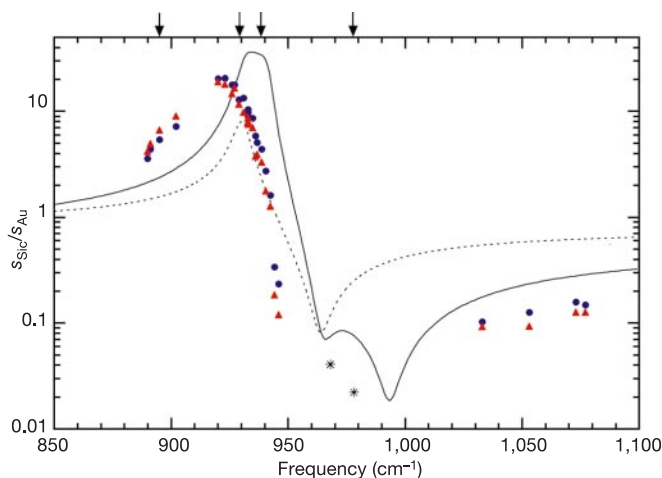


Figure 4 Phonon-enhanced near-field response of SiC, normalized to Au. The data points are extracted from near-field images such as those of Fig. 3, the two symbols corresponding to two different sample locations marked there by rectangles. At two frequencies (968 and 978 cm^{-1}) s_{SiC} is below the noise, which is marked instead by a black star. Arrows indicate the frequencies used for the images in Fig. 3c. The curves depict model predictions for the scattering amplitude s of a small Pt particle in contact with a SiC surface, normalized to an Au surface, equivalent to Fig. 2a (dashed line), and the same modified to simulate the experimental procedure of sinusoidal tapping motion, assuming $\Delta z = a$, and of second-harmonic signal demodulation (solid line).

1. Ebbesen, T. W., Lezec, H. J., Ghaemi, H. F., Thio, T. & Wolff, P. A. Extraordinary optical transmission through sub-wavelength hole arrays. *Nature* **391**, 667–669 (1998).
2. Xu, H., Bjerneld, E. J., Käll, M. & Börjesson, L. Spectroscopy of single hemoglobin molecules by surface enhanced Raman scattering. *Phys. Rev. Lett.* **83**, 4357–4360 (1999).
3. Specht, M., Pedarnig, J. D., Heckl, W. M. & Häscher, T. W. Scanning plasmon near-field microscopy. *Phys. Rev. Lett.* **68**, 476–479 (1992).
4. Zenhausern, F., Martin, Y. & Wickramasinghe, H. K. Scanning interferometric apertureless microscopy: optical imaging at 10 Angstrom resolution. *Science* **269**, 1083–1085 (1995).
5. Hillenbrand, R. & Keilmann, F. Material-specific mapping of metal/semiconductor/dielectric nanosystems at 10 nm resolution by back-scattering near-field optical microscopy. *Appl. Phys. Lett.* **80**, 25–27 (2002).
6. Knoll, B. & Keilmann, F. Near-field probing of vibrational absorption for chemical microscopy. *Nature* **399**, 134–137 (1999).
7. Barnes, W. L. Fluorescence near interfaces: the role of photonic mode density. *J. Mod. Opt.* **45**, 661–699 (1998).
8. Colas des Francs, G. *et al.* Optical analogy to electronic quantum corrals. *Phys. Rev. Lett.* **86**, 4950–4953 (2001).
9. Hillenbrand, R. & Keilmann, F. Optical oscillation modes of plasmon particles observed in direct space by phase-contrast near-field microscopy. *Appl. Phys. B* **73**, 239–243 (2001).
10. Joannopoulos, J. D., Villeneuve, P. R. & Fan, S. Photonic crystal: putting new twists on light. *Nature* **386**, 143–149 (1997).
11. Pendry, J. B. Negative refraction makes a perfect lens. *Phys. Rev. Lett.* **85**, 3966–3969 (2000).
12. Smith, D. R., Padilla, W. J., Vier, D. C., Nemat-Nasser, S. C. & Schultz, P. G. Composite medium with simultaneously negative permeability and permittivity. *Phys. Rev. Lett.* **84**, 4184–4187 (2000).
13. Krenn, J. R. *et al.* Squeezing the optical near-field zone by plasmon coupling of metallic nanoparticles. *Phys. Rev. Lett.* **82**, 2590–2593 (1999).
14. Weeber, J.-C. *et al.* Optical addressing on the subwavelength scale. *Phys. Rev. E* **62**, 7381–7388 (2000).
15. Ruppini, R. J. in *Electromagnetic Surface Modes* (ed. Boardman, A. D.) 345–398 (Wiley, Chichester, 1982).
16. Fröhlich, H. *Theory of Dielectrics* (Clarendon, Oxford, 1949).
17. Jackson, J. D. *Classical Electrodynamics* (Wiley & Sons, New York, 1975).
18. Palik, E. W. *Handbook of Optical Constants of Solids* (Academic, San Diego, 1985).
19. Engelbrecht, F. & Helbig, R. Effect of crystal anisotropy on the infrared reflectivity of 6H-SiC. *Phys. Rev. B* **48**, 15698–15707 (1993).
20. Kreibig, U., Gartz, M. & Hilger, A. Mie resonances: Sensors for physical and chemical cluster interface properties. *Ber. Bunsenges. Phys. Chem.* **101**, 1593–1604 (1997).
21. Faist, J. *et al.* Quantum cascade laser. *Science* **264**, 553–556 (1994).
22. Beck, M. *et al.* Continuous wave operation of a mid-infrared semiconductor laser at room temperature. *Science* **295**, 301–305 (2002).
23. Hillenbrand, R. & Keilmann, F. Complex optical constants on a subwavelength scale. *Phys. Rev. Lett.* **85**, 3029–3032 (2000).

24. Knoll, B. & Keilmann, F. Infrared conductivity mapping for nanoelectronics. *Appl. Phys. Lett.* **77**, 3980–3982 (2000).
25. Aravind, P. K. & Metiu, H. The effects of the interaction between resonances in the electromagnetic response of a sphere-plane structure; applications to surface enhanced spectroscopy. *Surf. Sci.* **124**, 506–528 (1983).
26. Shchegrov, A. V., Joulain, K., Carminati, R. & Greffet, J. J. Near-field spectral effects due to electromagnetic surface excitations. *Phys. Rev. Lett.* **85**, 1548–1551 (2000).
27. Hartmann, T., Kramer, A., Hillebrand, A., Guckenberger, R., et al. in *Procedures in Scanning Probe Microscopy* (ed. Colton, J. R.) 12–16 (Wiley, Chichester, 1998).
28. Aigouy, L. et al. Near-field optical spectroscopy using an incoherent light source. *Appl. Phys. Lett.* **76**, 397–399 (2000).

Acknowledgements

We acknowledge discussions with M. Stark, A. Otto, R. Helbig, R. Guckenberger and J. Plitzko. Supported by Deutsche Forschungsgemeinschaft.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.K. (e-mail: keilmann@biochem.mpg.de).

Pressure-induced insulator–conductor transition in a photoconducting organic liquid-crystal film

Chong-yang Liu & Allen J. Bard

Department of Chemistry and Biochemistry, The University of Texas at Austin, Austin, Texas 78712, USA

Intermolecular separation determines the extent of orbital overlap and thus the rate of electron transfer between neighbouring molecules in an organic crystal. If such a crystal is compressed, the resistivity decreases owing to a diminishing intermolecular distance¹. Metal–insulator transitions have been observed by applying hydrostatic pressure to, for example, Langmuir films of metal nanoparticles^{2,3}. But previous attempts to observe a clear transition point in organic crystals, such as anthracene and tetracene, were not successful owing to difficulties with electrically insulating the high-pressure cell⁴. Here we report a different approach by using a sample that is photoconductive and forms an organized film. A cylindrical tip (~100 μm in diameter) was used to compress the sample instead of a piston/cylinder structure, entirely eliminating the problem of electrical insulation. Furthermore, by illuminating the sample with a laser, the conductivity of the sample is increased by several orders of magnitude. By monitoring the photocurrent with sensitivity at the 10^{−13} A level, changes in resistivity at very low pressure could be monitored. We observe a sharp increase in current that could indicate a transition from hopping to delocalized conduction.

Zinc octakis(β-decoxyethyl)porphyrin (ZnODEP) was used as the sample. In the crystal, molecules are regularly stacked to form well-defined molecular columns. As the intercolumn separation is large (~24 Å), each column is fairly isolated electronically from a neighbouring column^{5,6}. In fact, in the liquid-crystal state (95–147 °C), the column structure still remains in the discotic mesophase, although the long tails are disordered⁵. The molecule-to-molecule distance within the same column is also quite large (~4 Å). The structure of ZnODEP suggests that the solid should be relatively soft and easy to compress compared to other organic

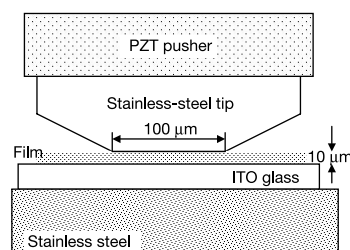


Figure 1 Schematic diagram of the experimental apparatus. A PZT pusher capable of moving extremely slowly (a few Å s^{−1}) was used to drive the tip so that the sample could be maintained in a quasi-equilibrium state during compression. A balance (Sartorius, 1212 MP) was used to measure the pressure precisely by placing the sample on its top.

crystals, such as anthracene. Films were made by a capillary filling of the substance in the molten state into empty ITO (indium tin oxide) sandwich cells as reported earlier^{7,8}. The cells were then cooled to form a solid film, and separated to produce a free surface. As shown in previous work^{7–9}, the axes of the molecular columns are oriented perpendicular to the ITO surface, so that the pressure would mainly affect the intermolecular interactions within the same columns.

An argon-ion laser (100 mW at 488 nm, Melles Griot) was used to irradiate the film via a fibre optic through the ITO glass. The tip was made from a stainless-steel cylinder about 0.1 inch in diameter with one end tapered to a sharp point by mechanical polishing. The tip was then pressed against a stainless-steel or glass plate to flatten the end to about 100–150 μm diameter by applying very high pressure. The overall structure is tip/film/ITO (Fig. 1). The electrical contacts were made to both the ITO and the tip electrodes. The tip was pushed by a PZT (lead zirconate titanate) pusher (Burleigh) to compress the film. A micrometer was attached to the other end of the pusher for the initial rough adjustment of the distance between the tip and the film.

When the tip was brought into contact with the ZnODEP film, a steady photocurrent was observed under a fixed bias voltage of 3.2 V (with the tip negative). The magnitude of the photocurrent depended on the initial film thickness. A short-circuit photocurrent could also be detected, as reported earlier, with large-area ITO/ZnODEP/ITO sandwich cells^{9–11}. This current is produced by preferential injection of photogenerated electrons from ZnODEP molecules into the illuminated ITO electrode, while holes hop through the film to the tip electrode. Note that the photocurrent started to increase immediately following the tip's forward move-

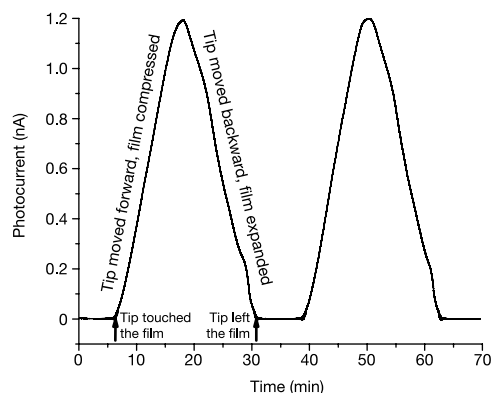


Figure 2 Photocurrent as a function of compression and expansion of a ZnODEP film about 6 μm thick between ITO glass and a stainless-steel tip. The ITO glass was placed on a fixed stainless-steel platform, and the tip was moved at a rate of 4 nm s^{−1}. The film was irradiated from the ITO side with a fibre optic through a hole in the platform. A bias voltage of 3.2 V was applied between sample and tip, with the tip negative.

ment to compress the film, and linearly increased from subpicoamp to nanoamp levels during the film compression. When the light was cut off briefly from time to time, the photocurrent dropped by several orders of magnitude, as expected, and increased back to the same level when the light was turned on again. When the film had been compressed to a certain level and the tip was then moved backward so that the film expanded, the photocurrent decreased. Reversible changes in photocurrent were observed during cycles of film compression and expansion (Fig. 2). The tip could be moved back and forth for many cycles at the same spot without affecting the reversibility and the magnitude of the photocurrent, indicating that the film beneath the tip was quite flexible and not physically damaged. Examination of the film under an optical microscope after the experiment showed no indentation or other damage at the position where the tip compression occurred, confirming that the applied pressure was mainly confined to an area over the molecular columns.

When the film was compressed further, a point was reached where the slope of the log current versus displacement curve greatly increased and the current became larger by several orders of magnitude (Fig. 3). A photocurrent at the $0.1 \mu\text{A}$ level was obtained, which was over six orders of magnitude higher than the initial value. This increase can be explained by a decrease in the intermolecular separation to such a level that the molecular orbitals overlap, changing the charge carrier transport mechanism from hopping to delocalized conduction. We do not know the pressure distribution along the molecular columns from the tip to the ITO substrate. Perhaps molecular separations near the tip were first decreased by compression at low pressure, while the intermolecular distances near the ITO surface were decreased at the critical pressure, leading to a very marked increase in current (Fig. 3). On the other hand, the molecular distances might be decreased relatively uniformly. In either case, however, the sharp current increase would signal a transition from hopping to delocalized conduction. We are not aware of any theory predicting such a transition.

The estimated resistivity of the ZnODEP film beneath the tip was of the order of $10^6 \Omega \text{ cm}$, indicating that the insulator was changed into a semiconductor. The pressure at the transition point was about 2.2 kbar, as determined with a balance (176 g over a tip about $100 \mu\text{m}$ in diameter). Seven measurements at different locations on the film showed the same result. Careful examination under an

optical microscope indicated that the shape of the tip was not changed after these measurements. Once the material became a semiconductor under pressure, the current was a function of bias voltage (Fig. 4). This current–voltage curve was very stable: ten consecutive curves essentially overlapped. The nonlinear behaviour clearly indicates that the tip was not in direct contact with the ITO substrate. In fact, the compressed film under the tip was still about $6 \mu\text{m}$ thick.

The same general behaviour was found for the film in the dark (Fig. 3). The initial current was over two orders of magnitude smaller, but a transition to a more conductive state was seen at essentially the same pressure as that for the illuminated film. Moreover, the optical properties of the ZnODEP film under pressure also changed when the compression was carried out on the stage of an inverted optical microscope (Nikon, TE300). The red colour of the film became darker with increasing pressure. Additional studies of the optical effects of compression are planned.

To determine how much the films were compressed under the tip, the samples were placed on a fixed stainless-steel platform instead of the balance. Before each measurement, the tip was directed first to the bare ITO surface next to the ZnODEP film. Once a tunnelling current between the tip and ITO was detected, the tip was withdrawn a known distance. The tip was then moved to approach the film and start the compression while the current and travel distance of the tip were monitored continuously. After the measurement, tip was moved to bare ITO surface again to assess the distance. This measured distance was exactly the same, even after the sample was taken out and put back onto the platform a number of times. Thus, the changes in the film thickness during the compression could be calculated. There is some uncertainty, however, because other parts, including the ITO glass, could be compressed to some degree. Because the glass and stainless steel are much harder than the ZnODEP film, their compression is probably negligible. The data indicate that the film was compressed by about 50% at the transition point. This can be compared to the 30–40% shrinkage observed with aromatic crystals, such as pentacene, at a pressure in the region of several hundred kilobars¹².

As mentioned above, the ZnODEP material is relatively soft owing to its structure (Fig. 3 inset), and is very different from conventional organic molecular crystals. Indeed, when anthracene films ($\sim 10 \mu\text{m}$ thick) were tested with our apparatus, the transition point was not seen under a pressure beyond our balance's upper

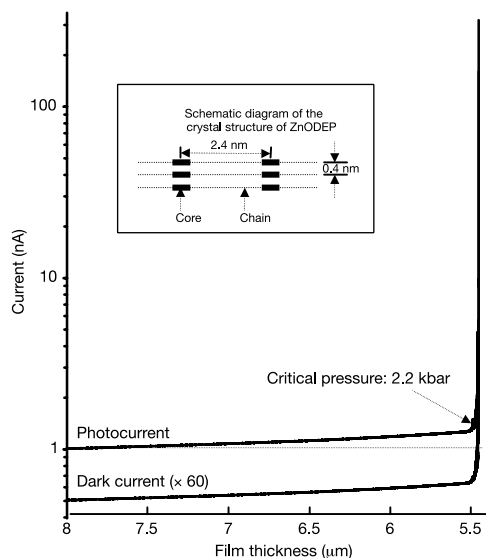


Figure 3 Photocurrent as a function of distance during the compression of a ZnODEP film about $10 \mu\text{m}$ thick between ITO glass and a stainless-steel tip. A bias voltage of 3.2 V was applied to the sample with the tip negative, and the film was compressed at a rate of 4 nm s^{-1} . Inset, schematic structure of ZnODEP.

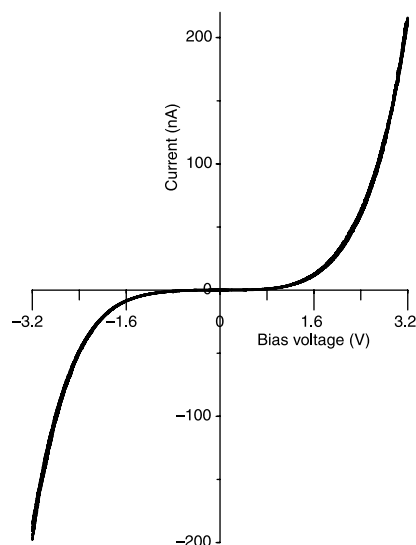


Figure 4 Current as a function of bias voltage in ZnODEP semiconductor (pressure about 2.7 kbar). Ten consecutive curves overlapped. Scan rate, 50 mV s^{-1} . The photocurrent was at 10^{-13} A under a bias voltage of 3.2 V when there was no compression.

limit, which was twice as high as the transition pressure for ZnODEP. □

Received 12 February; accepted 27 May 2002; doi:10.1038/nature00875.

- Drickamer, H. G. & Frank, C. W. Electronic structure, electronic transitions, and the high pressure chemistry and physics of solids. *Annu. Rev. Phys. Chem.* **23**, 39–64 (1972).
- Collier, C. P., Saykally, R. J., Shiang, J. J., Henrichs, S. E. & Heath, J. R. Reversible tuning of silver quantum dot monolayers through the metal-insulator transition. *Science* **277**, 1788–1791 (1997).
- Henrichs, S., Collier, C. P., Saykally, R. J., Shen, Y. R. & Heath, J. R. The dielectric function of silver nanoparticle Langmuir monolayers compressed through the metal insulator transition. *J. Am. Chem. Soc.* **122**, 4077–4083 (2000).
- Samara, G. A. & Drickamer, H. G. Effect of pressure on the resistance of fused-ring aromatic compounds. *J. Chem. Phys.* **37**, 474–479 (1962).
- Liu, C.-Y., Pan, H.-L., Tang, H., Fox, M. A. & Bard, A. J. Effect of structural order on the dark and photocurrents in zinc octakis (β-decoxyethyl)porphyrin thin-layer cells. *J. Phys. Chem.* **99**, 7632–7636 (1995).
- Schouten, P. G., Warman, J. M., de Haas, M. P., Fox, M. A. & Pan, H.-L. Charge migration in supramolecular stacks of peripherally substituted porphyrins. *Nature* **353**, 736–737 (1991).
- Liu, C.-Y. & Bard, A. J. Optoelectronic properties and memories based on organic single-crystal thin films. *Acc. Chem. Res.* **32**, 235–245 (1999).
- Liu, C.-Y. & Bard, A. J. in *Conjugated Oligomers, Polymers, and Dendrimers: From Polyacetylene to DNA* (ed. Brédas, J.-L.) 85–116 (De Boeck Université, Paris, 1999).
- Gregg, B. A., Fox, M. A. & Bard, A. J. Photovoltaic effect in symmetrical cells of a liquid crystal porphyrin. *J. Phys. Chem.* **94**, 1586–1598 (1990).
- Liu, C.-Y. & Bard, A. J. Optoelectric charge trapping/detrapping in thin solid films of organic azo dyes: Application of scanning tunneling microscopic tip contact to photoconductive films for data storage. *Chem. Mater.* **10**, 840–846 (1998).
- Liu, C.-Y., Pan, H.-L., Fox, M. A. & Bard, A. J. Reversible charge trapping/detrapping in a photoconductive insulator of liquid crystal zinc porphyrin. *Chem. Mater.* **9**, 1422–1429 (1997).
- Aust, R. B., Bentley, W. H. & Drickamer, H. G. Behavior of fused-ring aromatic hydrocarbons at very high pressure. *J. Chem. Phys.* **41**, 1856 (1964).

Acknowledgements

We thank F.-R.F. Fan for discussions. This work was supported by the National Science Foundation and the Robert A. Welch Foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.J.B. (e-mail: ajbard@mail.utexas.edu).

Self-regeneration of a Pd-perovskite catalyst for automotive emissions control

Y. Nishihata*, J. Mizuki*, T. Akao*†, H. Tanaka‡, M. Uenishi‡, M. Kimura§, T. Okamoto‡§ & N. Hamada||

* Synchrotron Radiation Research Center, Japan Atomic Energy Research Institute, Mikazuki, Sayo-gun, Hyogo 679-5148, Japan

‡ Materials R&D Division, Technical Center, Daihatsu Motor Co. Ltd, Ryuo, Gamo-gun, Shiga 520-2593, Japan

§ Materials Analysis & Evaluation Division, Toyota Central R&D Laboratories Inc., Nagakute, Aichi-gun, Aichi 480-1192, Japan

|| Faculty of Science and Technology, Tokyo University of Science, Noda, Chiba 278-8510, Japan

Catalytic converters are widely used to reduce the amounts of nitrogen oxides, carbon monoxide and unburned hydrocarbons in automotive emissions. The catalysts are finely divided precious-metal particles dispersed on a solid support. During vehicle use, the converter is exposed to heat, which causes the metal particles to agglomerate and grow, and their overall surface area to decrease. As a result, catalyst activity deteriorates. The

problem has been exacerbated in recent years by the trend to install catalytic converters closer to the engine, which ensures immediate activation of the catalyst on engine start-up, but also places demanding requirements on the catalyst's heat resistance. Conventional catalyst systems thus incorporate a sufficient excess of precious metal to guarantee continuous catalytic activity for vehicle use over 50,000 miles (80,000 km). Here we use X-ray diffraction and absorption to show that $\text{LaFe}_{0.57}\text{Co}_{0.38}\text{Pd}_{0.05}\text{O}_3$, one of the perovskite-based catalysts investigated^{1–4} for catalytic converter applications since the early 1970s, retains its high metal dispersion owing to structural responses to the fluctuations in exhaust-gas composition that occur in state-of-the-art petrol engines⁵. We find that as the catalyst is cycled between oxidative and reductive atmospheres typically encountered in exhaust gas, palladium (Pd) reversibly moves into and out of the perovskite lattice. This movement appears to suppress the growth of metallic Pd particles, and hence explains the retention of high catalytic activity during long-term use and ageing.

A state-of-the-art automotive petrol engine is operated close to the stoichiometric air-to-fuel ratio (by using an oxygen sensor and a sophisticated feedback control system linked to the catalyst) in order to convert simultaneously three pollutant emissions—carbon monoxide (CO), hydrocarbons, and nitrogen oxides (NO_x)⁵—into carbon dioxide (CO_2), water (H_2O) and nitrogen (N_2). A time lag associated with adjusting the air-to-fuel ratio results in a redox fluctuation in the exhaust gas. We designed our catalyst system to react to this fluctuation, to achieve greater efficiency and to conserve precious metals.

The conversion of CO and NO_x is equal at the CO– NO_x cross-over point (see Methods for catalytic evaluation), and the conversion at this point is generally accepted as a useful indicator of catalytic activity⁶. Figure 1 compares the activity of two catalysts using this indicator; one is $\text{LaFe}_{0.57}\text{Co}_{0.38}\text{Pd}_{0.05}\text{O}_3$ (Pd-perovskite, our 'intelligent' catalyst), and the other is Pd-impregnated $\gamma\text{-Al}_2\text{O}_3$ (Pd/alumina, the conventional catalyst; see Methods for catalyst preparation and testing). The Pd-perovskite catalyst maintains its high activity during ageing, whereas the activity of the Pd/alumina catalyst deteriorates by about 10%. (See Supplementary Information Fig. 1 for details on the conversion of different pollutants during the sweep test.) Imaging of the aged catalysts by transmission electron microscopy showed that the Pd particles on alumina reached sizes of up to 120 nm, whereas only small metallic particles about 1–3 nm in diameter were found on the perovskite surface (see Supplementary Information Fig. 2). X-ray energy dispersion analysis indicated that the small particles contain not only Pd, but also

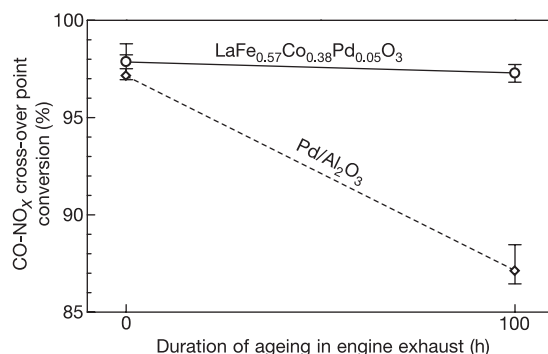


Figure 1 Change in catalyst activity during ageing. Shown is the ageing dependence of the CO– NO_x cross-over point conversion for $\text{LaFe}_{0.57}\text{Co}_{0.38}\text{Pd}_{0.05}\text{O}_3$ (Pd-perovskite catalyst) and Pd-impregnated $\gamma\text{-Al}_2\text{O}_3$ (Pd/alumina catalyst). The conversion efficiency was evaluated three times for each sample by the sweep test (see Methods). The median of the CO– NO_x cross-over point conversion (see Methods) is plotted as the symbol, together with the maximum and the minimum as error bars.

† Present addresses: Faculty of Engineering, Tottori University, Tottori 680-8552, Japan (T.A.); Research Institute for Science and Technology, Chubu University, Kasugai, Aichi 487-8501, Japan (T.O.).

Co. The remarkable suppression of metallic particle growth directly affects the capability of the Pd-perovskite catalyst to maintain high catalytic activity.

To investigate why the growth of Pd particles is suppressed on the Pd-perovskite catalyst, we conducted X-ray anomalous diffraction (XAD)⁷ and X-ray absorption fine structure (XAFS)⁸ measurements on oxidized, reduced and re-oxidized catalysts. The measurements were not conducted *in situ* or on samples treated in the presence of petrol exhaust; instead, the samples were simply oxidized in air, reduced by treatment in a hydrogen/nitrogen mixture, and re-oxidized in air. We expect that the structural changes seen in these samples will be similar to those occurring in the catalysts exposed to exhaust emissions in principle, but some differences may exist.

Figure 2a shows the powder diffraction pattern for the three $\text{LaFe}_{0.57}\text{Co}_{0.38}\text{Pd}_{0.05}\text{O}_3$ samples, for the limited range of momentum transfer Q that allows us to infer structural changes. Two Bragg peaks from the perovskite catalyst are assigned to the pseudocubic cell of the perovskite structure ABO_3 , where A and B cations are in 12-fold and 6-fold coordination, respectively. The positions of the Bragg reflections (100) and (110) for the reduced catalyst occur at lower angles than do the corresponding reflections for the oxidized catalyst, indicating expansion of the crystal lattice. Five additional peaks appeared in the spectrum of the reduced catalyst, which were assigned to lanthanum oxide (La_2O_3) and lanthanum hydroxide ($\text{La}(\text{OH})_3$). The spectrum indicates that the perovskite crystal structure is predominant, and that the material has partially transformed into $\text{La}(\text{OH})_3$ through La_2O_3 . The fraction of transformed material is difficult to estimate, but the X-ray absorption near edge structure (XANES) spectra at the Co K-edge imply that about half the Co, and thus about 20% of the perovskite, was transformed (see Supplementary Information Fig. 3). This quantity should depend on the reduction conditions. In the spectrum of the re-oxidized catalyst, the (100) and (110) reflections appear again at the same position as seen in the spectrum of the oxidized sample,

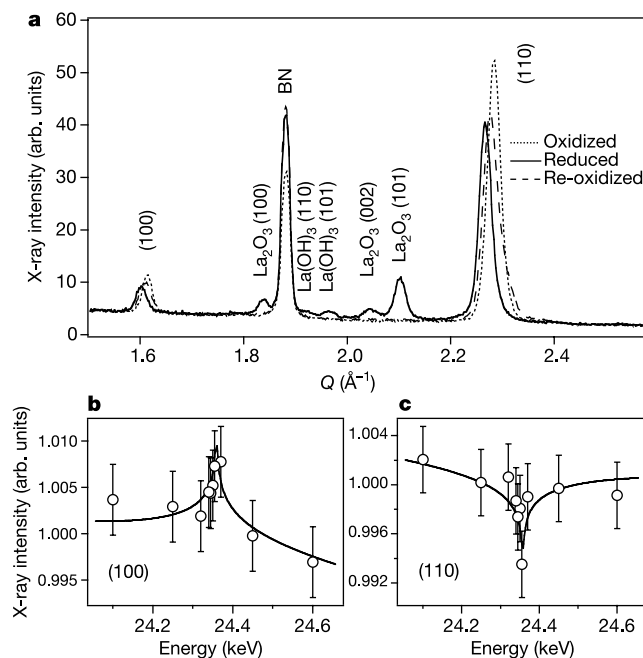


Figure 2 X-ray powder diffraction of $\text{LaFe}_{0.57}\text{Co}_{0.38}\text{Pd}_{0.05}\text{O}_3$. **a**, Diffraction pattern for catalyst samples oxidized in air, reduced in a hydrogen/nitrogen mixture and re-oxidized in air. The reflections of La_2O_3 and $\text{La}(\text{OH})_3$ appear only in the reduced catalyst. **b,c**, Energy dependence of the intensity for (100) (**b**) and (110) (**c**) reflections near the Pd K-edge indicates that the B-site of the perovskite structure is occupied by Pd in the oxidized catalyst.

while the extra peaks assigned to La_2O_3 and $\text{La}(\text{OH})_3$ have disappeared completely. These results indicate that the catalyst retains a predominantly perovskite structure throughout the oxidation, reduction and re-oxidation cycle, and that similar structural changes are likely to occur during the ageing experiment.

XAD was used to determine the distribution of Pd in the catalyst. The structure factors for the pseudocubic cell are expressed approximately in the form: $F(100) \propto |f_A - f_B - f_O|$ and $F(110) \propto |f_A + f_B - f_O|$, where f_A , f_B and f_O are atomic scattering factors for the A-site, B-site and oxygen atoms, respectively. Figure 2b and c represents the energy dependence of the reflection intensity for the oxidized catalyst. The cusp of the intensity of these reflections at the Pd K-edge indicates that Pd forms a solid solution with the perovskite crystal. Furthermore, by considering the structure factors involved, the increase of (100) and the decrease of (110) reflection intensities at the edge energy prove that Pd occupies the B-site. On the other hand, the cusp of the intensity was not observed for the reduced catalyst, implying that Pd does not occupy any site in the perovskite lattice of the reduced catalyst.

The valence state of Pd can be estimated from the XANES spectrum. Figure 3a shows the XANES spectra for Pd-perovskite samples after oxidation, reduction and re-oxidation, together with the XANES spectra of PdO and a Pd foil as reference materials. Compared to the PdO spectrum, the two oxidized samples show a chemical shift of about 1.7 eV, indicating that the valence of the Pd

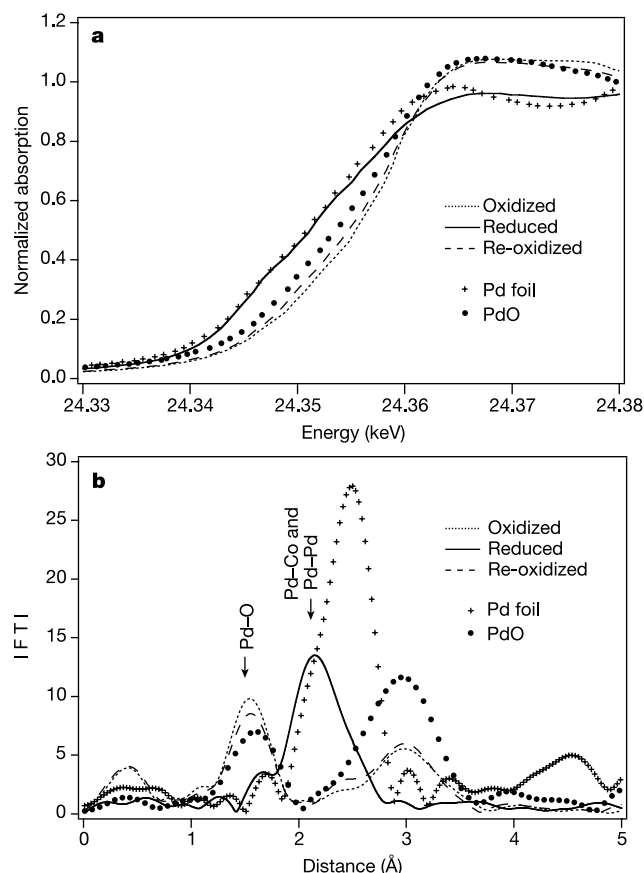


Figure 3 Comparison of XAFS data for $\text{LaFe}_{0.57}\text{Co}_{0.38}\text{Pd}_{0.05}\text{O}_3$, together with PdO and Pd foil as reference materials. **a**, XANES spectra were measured in transmission mode near the Pd K-edge. The valence state of Pd changes reversibly in a redox cycle. **b**, The radial structure function around Pd was calculated as the magnitude of the Fourier transform (FT) of the k^3 -weighted EXAFS oscillations. Generally, neighbours appear closer to the X-ray absorbing atom because the phase shift of the photoelectron was not taken into account. The first-nearest neighbour of Pd is indicated by an arrow for three catalysts. The local structure around Pd atom also changes reversibly.

Table 1 Local structure parameters around Pd estimated by EXAFS analysis

Sample	Shell	Coordination number	Interatomic distance (Å)	Debye–Waller factor (Å ²)	Discrepancy factor (%)
Oxidized catalyst	Pd–O	6*	2.023(4)	0.0046(2)	4.8
Reduced catalyst	Pd–Pd	5.2	2.661(2)	0.0069(1)	6.4
	Pd–Co	6.8			
Re-oxidized catalyst	Pd–O	6*	2.032(4)	0.0056(3)	6.0
PdO	Pd–O	4*	2.040(4)	0.0034(3)	6.4
Pd foil	Pd–Pd	12*	2.743(1)	0.0067(1)	4.2

The parameters marked with an asterisk were fixed. The sum of each shell coordination number for the reduced catalyst was fixed to 12 in consideration of a disordered f.c.c. structure. The coordination number of Pd–Pd bonding is larger than expected from the concentration of Pd and Co in the sample. This implies that all of the Co atoms did not segregate out from the perovskite crystal, or that Pd atoms associate preferentially with other Pd atoms. The discrepancy factor is defined as $\sum(k^3\chi_{\text{obs}} - k^3\chi_{\text{calc}})^2 / \sum(k^3\chi_{\text{obs}})^2$, where k , χ_{obs} and χ_{calc} are the wavenumber of a photoelectron excited from the X-ray absorbing atom, experimental and theoretical EXAFS functions, respectively.

in these materials is higher than the normal bivalence seen in PdO. The edge position of the reduced sample is identical to that of the Pd foil, suggesting that Pd is in a metallic state.

To elucidate the electronic state of Pd, we have performed first-principles electronic band structure calculations for Pd metal and an ordered La₂PdCoO₆ perovskite (the model material for our catalyst and based on LaCoO₃), by using the full-potential linearized augmented plane wave method in the local density approximation^{9,10}. The calculation shows that the empty Pd 5p density of states is lowered and broadened on going from La₂PdCoO₆ to Pd metal, which explains the change of the experimental XANES profile among the three catalysts. In La₂PdCoO₆, the Pd 4d *e_g* band is filled partially by just one electron; like Co, Pd at the B-site is therefore trivalent. Metal ions are known to be able to exist in such unusual oxidation states in the perovskite structure by ion substitution^{11–13}.

The radial structure function around Pd was calculated by Fourier transform of extended X-ray absorption fine structure (EXAFS) oscillations (Fig. 3b). The local structure parameters concerning the first nearest neighbour of Pd were evaluated by nonlinear least squares fitting (Table 1). The interatomic distance between Pd and O for the oxidized catalyst sample is close to half the lattice constant for the pseudocubic cell (1.94 Å). The ionic radii of Pd²⁺, Pd³⁺ and Pd⁴⁺ with six-fold coordination were reported to be 0.86 Å, 0.76 Å and 0.62 Å, respectively¹⁴. As the ionic radius of O^{2–} is 1.40 Å (ref. 14), trivalent or tetravalent Pd is expected from consideration of the interatomic distance in Table 1. This is also consistent with the observed chemical shift of the absorption edge for the oxidized sample. The XAFS data thus support our conclusion that Pd occupies the B-site in the oxidized sample. For the reduced catalyst, the XAD and XANES measurements suggested the segregation of metallic Pd from the perovskite crystal, while the segregation of Co was implied by X-ray energy dispersion analysis. XAFS data at the Co K-edge also indicated that about half the Co left the perovskite crystal lattice and existed in a metallic state, while the remainder kept occupying the B-site (Supplementary Information Fig. 3). The fitting model for the metallic material suggests a disordered face-centred cubic (f.c.c.) lattice containing Pd and Co, as Pd is known to form a solid solution with Co in the f.c.c. structure¹⁵ over a wide concentration range.

This model is essentially a two-shell model, consisting of Pd–Co and Pd–Pd bonds having the same length and thermal parameter (Debye–Waller factor). As seen from the interatomic distances (Table 1), the average length given by the two-shell model, which should fall between the Pd–Co and Pd–Pd bond lengths, is comparable with the sum of the atomic radii of Pd (1.376 Å) and Co (1.253 Å)¹⁶. The local structure of the re-oxidized sample is essentially the same as that of the oxidized sample, indicating that it can be changed in a completely reversible manner. This evidence for Pd movement suggests that Pd might also move back and forth between the B-site in the perovskite structure and the metal particle lattice site in the ‘real’ catalyst, when exposed to fluctuations in the redox characteristics of the emission exhaust. □

Methods

Preparation and rapid ageing of monolithic catalysts

Our Pd-perovskite catalyst, LaFe_{0.57}Co_{0.38}Pd_{0.05}O₃, was prepared by the alkoxide method^{17,18}. This involved dissolving metal ethoxyethylates M³⁺(OC₂H₄OC₂H₅)₃ (M = La, Fe and Co) in toluene. To prepare 1 mol of perovskite oxides, metal ethoxyethylates M³⁺(OC₂H₄OC₂H₅)₃ (M = La, Fe and Co) were dissolved in 200 cm³ of toluene. Precipitates containing Pd were obtained by hydrolysis with a diluted Pd(NO₃)₂ aqueous solution (50.89 mg cm^{–3} as Pd metal). After drying, the precursor was calcined at 700 °C for 4 h in air to obtain a perovskite powder catalyst containing Pd homogeneously. A Pd-impregnated γ-Al₂O₃ (Pd/alumina catalyst) was also prepared as a conventional standard. Each catalyst was coated on the inner walls of a monolithic honeycomb substrate (80 mm in diameter and 95 mm in length, with a grid of 64 cells per cm²), containing the same amounts of Pd (3.24 mg cm^{–3}). In order to investigate overall lifetime performance, the two monolithic catalysts were attached to each side bank in the exhaust system of an automotive 4,000-cm³ V8 engine for the simultaneous ageing experiment. Then the catalysts were exposed to the high-temperature exhaust gas at 900 °C for 100 h. During exposure, two exhaust gas conditions were alternated; an atmosphere with large redox fluctuations ($\Delta\lambda = \pm 4\%$ at 0.6 Hz; see below) around the stoichiometric point was applied for 870 s, then an oxidizing atmosphere for 30 s. This ageing procedure is accepted as appropriate for the simulation of a driving strategy to achieve higher fuel economy by introducing fuel cut-off during deceleration^{5,6}. The parameter λ is used to define the redox atmospheres: $\lambda = (\text{the present air-to-fuel ratio}) / (\text{the stoichiometric air-to-fuel ratio})$, where $\lambda = 1$ shows the stoichiometric air-to-fuel ratio, $\lambda < 1$ means a reducing atmosphere, and $\lambda > 1$ means an oxidizing atmosphere.

Evaluation of catalytic activity

The sweep test was carried out so as to evaluate conversion of the three pollutant emissions by shifting the mean value of λ from 0.890 to 1.096 with the fluctuation $\Delta\lambda = \pm 3.4\%$ at 0.5 Hz at 400 °C in a space velocity of 35,000 h^{–1}. Conversion of CO is improved in the lean region ($\lambda > 1$), whereas the NO_x reduction occurs readily in the rich region ($\lambda < 1$). The conversion is equal for CO and NO_x near the stoichiometric point ($\lambda = 1$), which is referred as the CO–NO_x cross-over point conversion. As a state-of-the-art automotive petrol engine is operated around the stoichiometric point within an operating window⁵, the CO–NO_x cross-over point conversion is an indicator for the activity of the automotive catalyst.

Thermal treatment of catalyst for X-ray measurements

This was carried out in three steps: the powdered catalyst was oxidized in air at 800 °C for 1 h, then reduced in an atmosphere of 10% H₂/90% N₂ at 800 °C for 1 h, and finally re-oxidized in air at 800 °C for 1 h. After each step, the sample was mixed with a binder of BN, and pressed to form into pellets. X-ray diffraction and XAFS measurements near the Pd K-edge (24.350 keV) were carried out at bending-magnet beamline BL14B1 of the 8-GeV synchrotron radiation source, SPring-8.

Received 5 November 2001; accepted 27 May 2002; doi:10.1038/nature00893.

- Meadowcroft, D. B. Low-cost oxygen electrode material. *Nature* **226**, 847–848 (1970).
- Libby, W. F. Promising catalyst for auto exhaust. *Science* **171**, 499–500 (1971).
- Tanaka, H., Fujikawa, H. & Takahashi, I. *Perovskite-Pd Three-Way Catalysts for Automotive Applications* (SAE Paper 930251, Society of Automotive Engineers, Inc., Warrendale, Pennsylvania, 1993).
- Tanaka, H., Fujikawa, H. & Takahashi, I. *Excellent Oxygen Storage Capacity of Perovskite-Pd Three-Way Catalysts* (SAE Paper 950256, Society of Automotive Engineers, Inc., Warrendale, Pennsylvania, 1995).
- Heck, R. M. & Farrauto, R. J. *Catalytic Air Pollution Control: Commercial Technology* 94–102 (Van Nostrand Reinhold, New York, 1995).
- Waltner, A. et al. *Development of Close-Coupled Catalyst Systems for European Driving Conditions* (SAE Paper 980663, Society of Automotive Engineers, Inc., Warrendale, Pennsylvania, 1998).
- Materlik, G., Sparks, C. J. & Fischer, K. (eds) *Resonant Anomalous X-Ray Scattering* (North-Holland, Amsterdam, 1994).
- Teo, B. K. *EXAFS: Basic Principles and Data Analysis* (Springer, Berlin, 1986).
- Vosko, S. H., Wilk, L. & Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **58**, 1200–1211 (1980).
- Painter, G. S. Improved correlation corrections to the local-spin-density approximation. *Phys. Rev. B* **24**, 4264–4270 (1981).
- Watanabe, H. Magnetic properties of perovskites containing strontium. *J. Phys. Soc. Jpn* **12**, 515–522 (1957).
- Jonker, G. H. Semiconducting properties of lanthanum-cobalt oxide. *Philips Res. Rep.* **24**, 1–14 (1969).
- Tejeda, L., Fierro, J. L. G. & Tascon, J. M. D. in *Advances in Catalysis* (eds Eley, D. D., Pines, H. & Weisz, P. B.) Vol. 36, 237–328 (Academic, San Diego, 1989).

14. Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A* **32**, 751–767 (1976).
15. Hansen, M. (ed.) *Constitution of Binary Alloys* (McGraw-Hill, New York, 1958).
16. Emsley, J. (ed.) *The Elements* (Clarendon, Oxford, 1994).
17. Tanaka, H., Takahashi, I., Kimura, M. & Sobukawa, H. in *Science and Technology in Catalysts 1994* (eds Izumi, Y., Arai, H. & Iwamoto, M.) 457–460 (Kodansha-Elsevier, Tokyo, 1994).
18. Tanaka, H., Tan, I., Uenishi, M., Kimura, M. & Dohmae, K. in *Topics in Catalysts* (eds Kruse, N., Frennet, A. & Bastin, J.-M.) Vols 16/17, 63–70 (Kluwer Academic, New York, 2001).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank M. Misono and N. Mizuno for their advice on catalyst research.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.N. (e-mail: yasuo@spring8.or.jp).

A satellite geodetic survey of large-scale deformation of volcanic centres in the central Andes

Matthew E. Pritchard & Mark Simons

Seismological Laboratory, Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA

Surface deformation in volcanic areas usually indicates movement of magma or hydrothermal fluids at depth. Stratovolcanoes tend to exhibit a complex relationship between deformation and eruptive behaviour¹. The characteristically long time spans between such eruptions requires a long time series of observations to determine whether deformation without an eruption is common at a given edifice. Such studies, however, are logistically difficult to carry out in most volcanic arcs, as these tend to be remote regions with large numbers of volcanoes (hundreds to even thousands). Here we present a satellite-based interferometric synthetic aperture radar (InSAR) survey of the remote central Andes volcanic arc, a region formed by subduction of the Nazca oceanic plate beneath continental South America. Spanning the years 1992 to 2000, our survey reveals the background level of activity of about 900 volcanoes, 50 of which have been classified as potentially active^{2,3}. We find four centres of broad (tens of kilometres wide), roughly axisymmetric surface deformation. None of these centres are at volcanoes currently classified as potentially active, although two lie within about 10 km of volcanoes with known activity. Source depths inferred from the patterns of deformation lie between 5 and 17 km. In contrast to the four new sources found, we do not observe any deformation associated with recent eruptions of Lascar, Chile^{4,5}.

We have created radar interferograms imaging about 900 of the approximately 1,100 catalogued volcanic edifices in the central Andes² (Fig. 1). Included in our study are several silicic calderas and geothermal fields. Temporal coverage is variable—for some locations, ERS 1 and 2 (European Space Agency remote-sensing satellites) radar data are available from 1992 to 2000, but only four years of data are available for a few volcanoes.

We infer four major sources of deformation from the InSAR data. We find that Uturuncu, a stratovolcano in Bolivia, has a maximum deformation rate in the radar line-of-sight (LOS) direction, U_{LOS} ,

of 1–1.5 cm yr^{−1} (Fig. 2). An area in southern Peru about 2.5 km east of the volcano Hualca Hualca and 7 km north of the active volcano Sabancaya is inflating with U_{LOS} of about 2 cm yr^{−1}. A third inflationary source (with $U_{\text{LOS}} = 1$ cm yr^{−1}) is not associated with a volcanic edifice. This third source is located 11.5 km south of Lastarria and 6.8 km north of Cordon del Azufre on the border between Chile and Argentina, and is hereafter called 'Lazufre'. Robledo caldera, in northwest Argentina, is subsiding with U_{LOS} of 2–2.5 cm yr^{−1}. Because the inferred sources are more than a few kilometres deep, any complexities in the source region are damped such that the observed surface deformation pattern is smooth. We find no measurable deformation at volcanoes with small eruptions or fumarolic activity during the period when radar observations were made—Ubinas (in Peru) or Guallatiri, Irruputuncu and Ojos del Salado (in Chile)⁶.

It is possible that other volcanoes are deforming at rates below our detection threshold. We quantify our sensitivity to deformation by comparing temporally overlapping interferograms, and by consideration of previous analyses of InSAR accuracy⁷. We estimate that we can detect a signal of 1–2 cm amplitude that is at least 10 km in spatial extent. Differentiating such a signal from atmospheric noise within a single interferogram is difficult, but is possible by comparing multiple interferograms, as well as multiple edifices within an interferogram. Owing to decorrelation of the radar signal near many of the volcano summits, our survey is not very sensitive to localized near-surface deformation, as seen for example in the Galapagos⁸. Because the longest interferograms that we can form for most regions span about 5 yr (limited by data availability and maintaining interferometric coherence), we can detect average deformation rates exceeding 4 mm yr^{−1}.

We model the observed surface deformation as resulting from a spherical point-source of volume change, embedded in either a homogeneous or layered elastic crust. For each volcano, we simultaneously invert data from as many different time periods and satellite tracks as possible, assuming a single source location. We solve for the volume change in each interferogram separately. For our inversions, we use the Neighbourhood Algorithm global search approach, which can find multiple local minima and a range of acceptable models⁹. The misfit function is usually peaked near the best estimate, but because of the data noise and non-uniqueness of the problem, we instead choose to use the width of the misfit function to estimate a range of acceptable values.

We consider how three different seismically derived one-dimensional layered elastic models^{10,11} affect estimates of source parameters. We incorporate the effects of topography upon source depth by using a different source depth for each pixel depending upon local elevation¹². Our experiments with different elastic structures, topography, and the trade-off between source depth and volume change indicate a range of possible source depths for each centre of deformation spanning a few kilometres.

From north to south, the inferred source depths (below sea level) are: 11–13 km at Hualca Hualca; 15–17 km for Uturuncu; 6–8 km for Lazufre; and 4.5–6 km at Robledo. Thus, we find no obvious pattern in source depth within this volcanic arc. The fit of the models to the data is generally good (root mean square, r.m.s. < 1 cm, Fig. 2), although a consistent residual northeast of Hualca Hualca might indicate localized subsidence, and the oblong shape of the Lazufre deformation pattern suggests the need for a non-spherical source.

The rate of volume change seems to be time-dependent at Uturuncu, Lazufre and Robledo, whereas the deformation rate at Hualca Hualca seems constant within measurement error (Fig. 3). At Lazufre, no deformation is apparent before early 1998 in two averaged interferograms, but there is a clear signal in three later interferograms. The temporal coverage is insufficient to resolve whether the start of deformation was abrupt or gradual. At Uturuncu, there is also a slight increase in inflation rate after early

1998. It could be coincidental that the two centres have an increased rate of inflation at roughly the same time, but we note that a $M_w = 7$ subduction zone earthquake occurred near the time of the increase (30 January 1998, Fig. 1). Correlations between unrest at volcanic centres and regional earthquakes have been documented before¹³. At Robledo, the rate of subsidence seems to be decreasing with time.

Even at well-studied volcanoes, the cause of the deformation is often ambiguous. Given the paucity of relevant studies in the central

Andes, we emphasize that the cause of deformation we observe is not known. Inflation at Hualca Hualca, Uturuncu and Lazufre could be caused by injection of magma from depth, melting of crustal rock from a previous injection, or build-up of pressure from a hydrothermal system. Given the inferred 15–17 km source depth for deformation at Uturuncu, the source is probably magmatic, not hydrothermal in nature. This source is part of the Altiplano–Puna Magmatic Complex¹⁴, and may be related to a region of low seismic velocity inferred to indicate a zone of partial melt at 19 km (ref. 15).

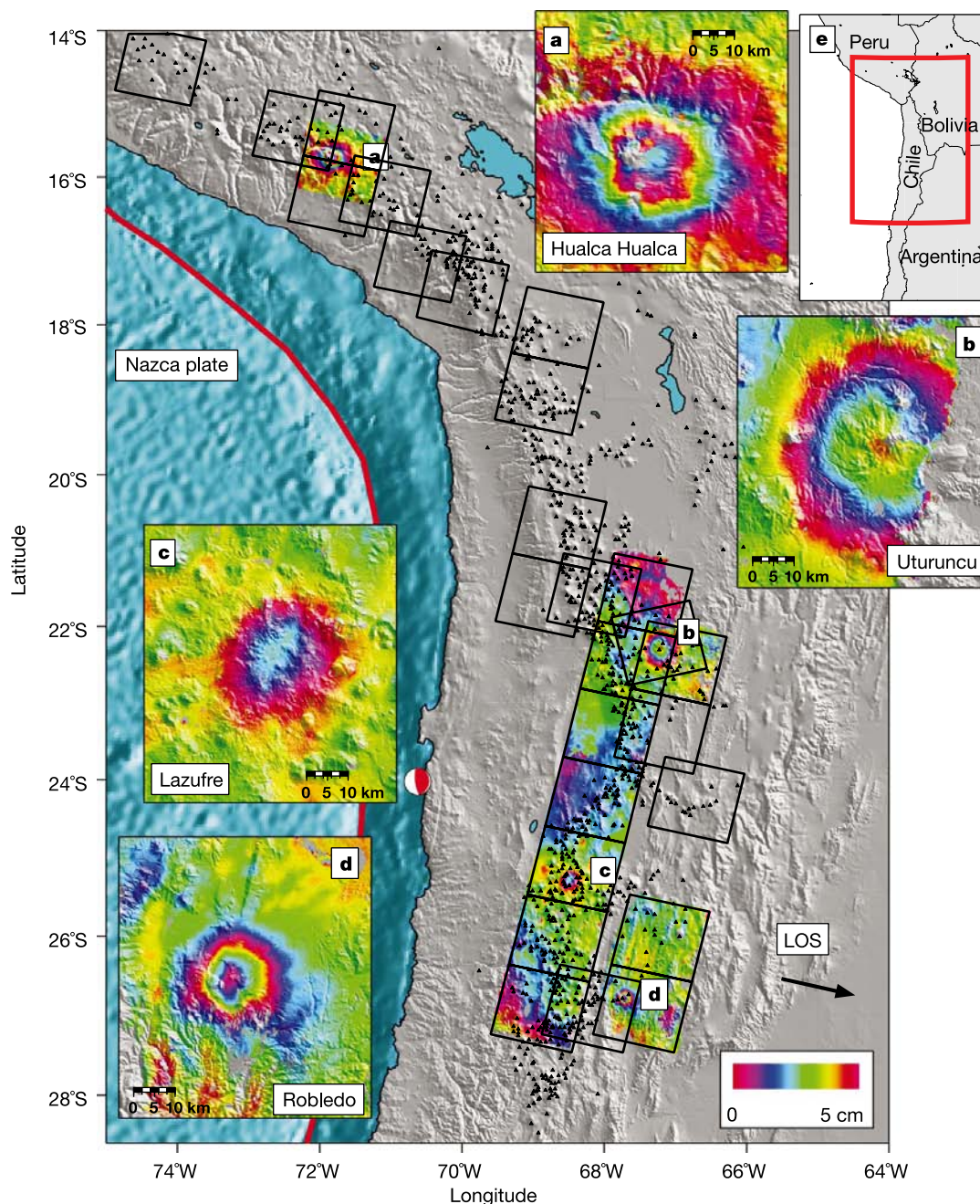


Figure 1 Shaded relief topographic map of the central Andes. The radar frames used in this study are shown (black squares). Black triangles show the 1,113 potential volcanic edifices². The red line in the ocean is the location of the trench. Colour shows interferograms (draped over shaded relief) from this study indicating active deformation—each colour cycle corresponds to 5 cm of deformation in the radar line-of-sight (LOS) direction. Arrow shows radar LOS direction from ground to spacecraft, which is inclined 23° from the vertical. The red and white circle indicates the location of the $M_w = 7$

subduction zone earthquake mentioned in the text that is temporally correlated with increases in the rates of inflation at Uturuncu and Lazufre. Inset maps provide detailed views of the centres of volcanic deformation: **a**, Hualca Hualca, Peru, time span June 1992–April 1996 (3.9 yr); **b**, Uturuncu, Bolivia, time span May 1996–December 2000 (4.6 yr); **c**, Lazufre, time span May 1996–December 2000 (4.6 yr), all of which are inflating; and **d**, Robledo, time span May 1996–October 2000 (4.4 yr), in northwest Argentina, which is deflating. **e**, Red box places main figure in context.

Lazufre and Robledo also lie near regions with low seismic velocities¹⁰. The deformation source near Hualca Hualca could be related to activity at Sabancaya, even though it is 7 km away. During the 1990 eruptions of Sabancaya, a concentration of earthquakes was located between 4–7 km depth in about the same region as our inferred deformation source¹⁶. A large distance between an eruptive centre and the magma chamber has been inferred elsewhere; for example, in Alaska, where the 1912 Novarupta eruption was at least partially fed by a magma chamber beneath Mt Katmai 10 km distant¹⁷. Lazufre also lies more than 7 km from the nearest potentially active volcano. This inferred deformation source could plausibly feed either of the nearby volcanoes, or—as it lies within a topographic depression—it could represent an incipient caldera. Subsidence at the southernmost source, Robledo, is most probably caused by cooling or crystallization of a magma body or hydrothermal activity, and not by lateral draining of magma, because no deformation signal from such magma movement is seen.

Lascar, Chile, is currently the most active volcano in the central Andes and has had several eruptions during the period when InSAR data are available. We do not observe any deformation at Lascar in interferograms spanning the eruptions in April 1993, December 1993, and July 2000, or those spanning other periods of mild vulcanian and fumarolic activity (Fig. 4). In the co-eruptive interferograms, there is a small region of interferometric decorrelation around the edifice, but even so, the large expected deformation signal from the estimated volume of material erupted cannot hide at shallow depths. However, if the source is deep, the corresponding surface deformation signal could be below our detection threshold, thereby placing a constraint on the minimum chamber depth. Assuming a spherical source in an elastic half space and a range in volume of erupted material (10^8 m^3 for April 1993; 10^6 – 10^7 m^3 for December 1993 and July 2000)¹⁸, the minimum depths below local relief for the April 1993 eruption is 20–25 km, and 5–12 km for the December 1993 and July 2000 eruptions. Petrological estimates of the magma chamber depth range from 5–6 km (ref. 19) to 12–22 km

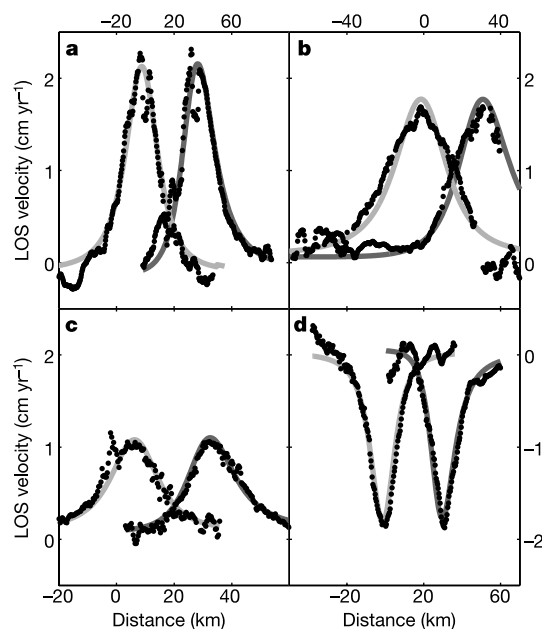


Figure 2 South to north (light grey) and west to east (dark grey) profiles through the interferometric observations (points, same data as shown in Fig. 1 in map insets and model fits (lines), both shown as rates. The west–east profile has been offset for the sake of clarity. The profiles intersect at the centre of their respective deformation patterns. **a**, Hualca Hualca. In these profiles, the data points are consistently below the models owing to a localized region of subsidence to the northeast of Hualca Hualca that is visible in Fig. 1. **b**, Uturuncu. **c**, Lazufre. **d**, Robledo.

(ref. 20). Alternatively, the absence of surface deformation could indicate that the magma plumbing system (particularly the vertical conduit) behaved rigidly and did not deform during the eruption. We note that gravity measurements at several other volcanoes appear to indicate magma movements without surface deformation^{21–23}. Finally, the lack of surface deformation may result from temporally aliased observations related to the cyclic eruptive style of Lascar. At Lascar, a lava dome grows within the crater, cools and subsides, inhibiting degassing and building pressure within the chamber until eruption⁴. In each case, our co-eruptive interferograms begin before the subsidence and pressure build-up commenced, so that pre-eruptive inflation of Lascar could have been approximately balanced by post-eruptive collapse, giving little net deformation in the interferogram. All three alternatives are viable, with further work required before one explanation can be favoured.

Our survey suggests the need for monitoring and hazard assessment near the four sources of deformation. Potential hazards from eruption include mudflows (triggered by melting snow), landslides, pyroclastic flows and ash clouds, which could endanger overflying aircraft. The volcanoes of southern Peru pose the most significant

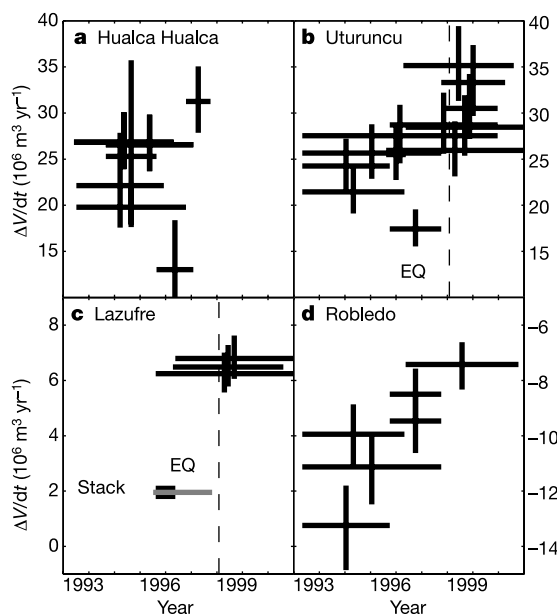


Figure 3 Inferred rate of volume change as a function of time, assuming a constant source depth at each location, a spherically symmetric source in a half-space, and a constant rate of deformation during the time period covered in each interferogram. The horizontal bar shows the time period covered by the interferogram and the vertical bar reflects an estimate of the error on the inferred rate of volume change. The error bar is 0.05 units in log space (such that the absolute error scales with the size of the source), except for scenes with extensive atmospheric contamination at Hualca Hualca, where we estimate the error to be 0.10 log units. The vertical error bar has been estimated by looking at the spread in the scatter plot of misfit as a function of source strength, comparing the strength results from inversions of different combinations of data sets, and comparing interferograms that span nearly the same time interval, including a set of interferograms at Robledo that differ by only one day. **a**, Hualca Hualca. **b**, Uturuncu shows an increase in inflation following a $M_w = 7$ Chilean subduction zone earthquake. The time of the earthquake is shown by a dashed vertical line (labelled EQ). **c**, Lazufre. Two interferograms before 1998 have been averaged (the black and grey horizontal bars labelled Stack represent the time span of the two interferograms) and show no deformation. The stacked data place an upper bound on the rate of source volume change. Three interferograms that include times following the $M_w = 7$ Chilean subduction zone earthquake show deformation. **d**, Robledo. The total volume of material inferred to have moved during the observation period for each source is: Uturuncu, $2.4 \times 10^9 \text{ m}^3$; Robledo, $-9 \times 10^7 \text{ m}^3$; Hualca Hualca, $1.3 \times 10^8 \text{ m}^3$; and Lazufre, $3.5 \times 10^7 \text{ m}^3$.

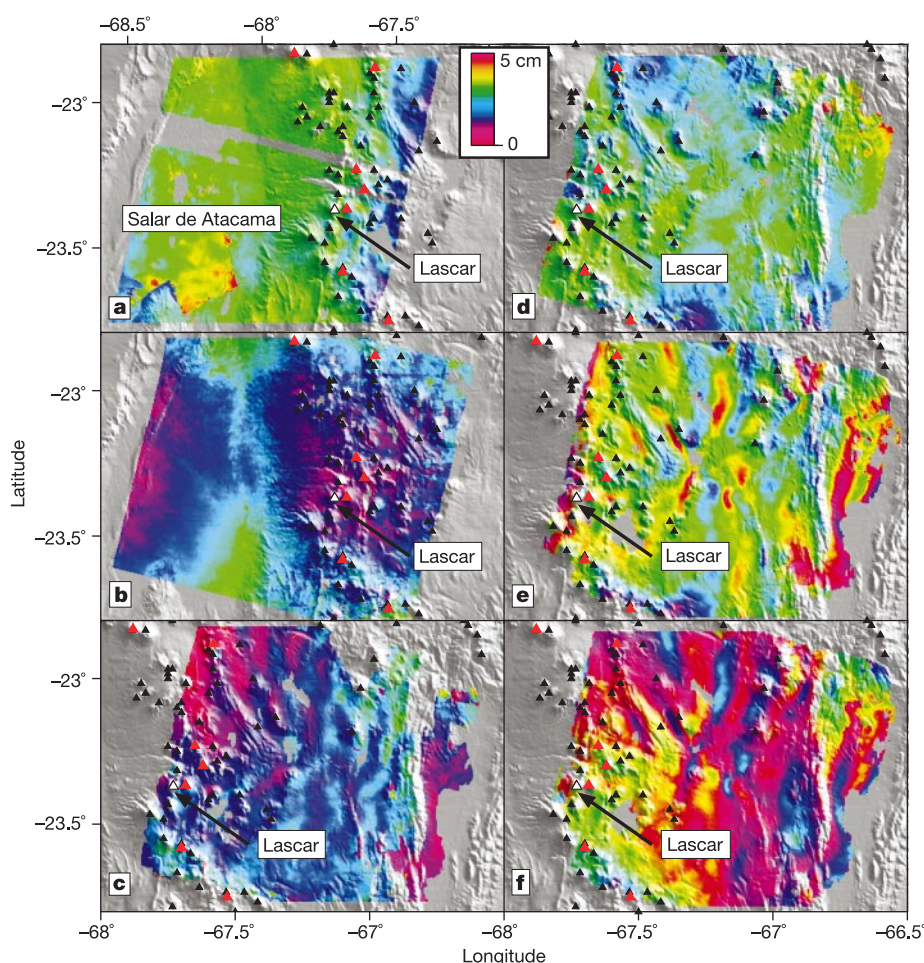


Figure 4 Interferograms showing no deformation at Lascar from two satellite tracks of radar data. Black triangles show volcanic edifices, red triangles are potentially active volcanoes, and the white triangle is Lascar. **a**, 12 August 1995–24 December 2000 (track 282), spanning the 20–21 July 2000, eruption and a quiet vapour emission⁶. Deformation in the Salar de Atacama is clearly visible. **b**, 6 August–19 March 2000 (track 282), spanning the eruption on 20–21 July 2000 (refs 5, 6). **c**, 26 March 1996–13 November 1993 (track 10), spanning the eruption on 17 December 1993 (refs 4, 6, 25), and small

eruptions March 1994–June 1995 (ref. 6). **d**, 7 October 1997–2 October 1995 (track 10), during which the only documented activity was a quiet vapour emission⁶. **e**, 2 October 1995–2 May 1992 (track 10) spanning the major eruption on 18–20 April 1993, the eruption on 17 December 1993 (refs 3, 4, 6, 25), and other small eruptions March 1994–June 1995 (ref. 6). **f**, 2 October 1995–2 May 1992 (track 10), spanning the major eruption on 18–20 April 1993 the eruption on 17 December 1993 (refs 3, 4, 6, 25), small eruptions March 1994–June 1995 (ref. 6), and a quiet vapour emission⁶.

hazard because of the higher population density in that area. For example, an eruption from Hualca Hualca, Sabancaya or Ampato could endanger the 35,000 people who live near this group of volcanoes²⁴. □

Received 30 April; accepted 27 May 2002; doi:10.1038/nature00872.

1. Dvorak, J. J. & Dzursin, D. Volcano geodesy: The search for magma reservoirs and the formation of eruptive vents. *Rev. Geophys.* **35**, 343–384 (1997).
2. de Silva, S. L. & Francis, P. W. *Volcanoes of the Central Andes* (Springer, New York, 1991).
3. Gonzalez-Ferran, O. *Volcanes de Chile* (in Spanish) (Instituto Geografico Militar, Santiago, Chile, 1995).
4. Matthews, S. J., Gardeweg, M. C. & Sparks, R. S. J. The 1984 to 1996 cyclic activity of Lascar Volcano, northern Chile: Cycles of dome growth, dome subsidence, and explosive eruptions. *Bull. Volcanol.* **59**, 77–82 (1997).
5. Wooster, M. J. Long-term infrared surveillance of Lascar Volcano: Contrasting activity cycles and cooling pyroclastics. *Geophys. Res. Lett.* **28**, 847–850 (2001).
6. Smithsonian Institution Global Volcanism Program. (<http://nmnhwww.si.edu/gvp/>) (accessed 29 April 2002).
7. Hanssen, R. A. *Radar Interferometry: Data Interpretation and Error Analysis* (Kluwer Academic, Dordrecht, The Netherlands, 2001).
8. Amelung, F., Jonsson, S., Zebker, H. & Segall, P. Widespread uplift and 'trapdoor' faulting on Galapagos volcanoes observed with radar interferometry. *Nature* **407**, 993–996 (2000).
9. Sambridge, M. Exploring multi-dimensional landscapes without a map. *Inverse Problems* **14**, 427–440 (1998).
10. Yuan, X. *et al.* Subduction and collision processes in the Central Andes constrained by converted seismic phases. *Nature* **408**, 958–961 (2000).

11. Wigger, P. J. *et al.* in *Tectonics of the Southern Central Andes: Structure and Evolution of an Active Continental Margin* (eds Reutter, K.-J., Scheuber, E. & Wigger, P. J.) 23–48 (Springer, Berlin, 1994).
12. Williams, C. A. & Wadge, G. The effects of topography on magma chamber deformation models: Application to Mt. Etna and radar interferometry. *Geophys. Res. Lett.* **25**, 1549–1552 (1998).
13. Johnston, M. J. S., Hill, D. P., Linde, A. T., Langbein, J. & Bilham, R. Transient deformation during triggered seismicity from the 28 June Mw = 7.3 Landers earthquake at Long-Valley caldera, California. *Bull. Seismol. Soc. Am.* **85**, 787–795 (1995).
14. de Silva, S. L. Altiplano-Puna volcanic complex of the central Andes. *Geology* **17**, 1102–1106 (1989).
15. Chmielowski, J., Zandt, G. & Haberland, C. The central Andean Altiplano-Puna magma body. *Geophys. Res. Lett.* **26**, 783–786 (1999).
16. Lazo, M., Kosaka, A., Minaya, A., Gonzales, E. & Soto, J. in *Proc. VII Congreso Peruano de Geología* 19–22 (Sociedad Geologica del Peru, Lima, 1991).
17. Curtis, G. H. in *Studies in Volcanology, a Memoir in Honor of Howell Williams* Geol. Soc. Mem. 116 (eds Coats, R. R., Hay, R. L. & Anderson, C. A.) 153–210 (GSA, Boulder, CO, 1968).
18. Simkin, T. & Siebert, L. *Volcanoes of the World* (Geoscience, Tucson, 1994).
19. Matthews, S. J., Sparks, R. S. J. & Gardeweg, M. C. Piedras Grandes-Soncor eruptions, Lascar volcano, Chile; evolution of zoned magma chamber in the central Andean upper crust. *J. Petrol.* **40**, 1891–1919 (1999).
20. Matthews, S. J., Jones, A. P. & Gardeweg, M. C. Lascar volcano, northern Chile—Evidence for steady-state disequilibrium. *J. Petrol.* **35**, 401–432 (1994).
21. Rymer, H., Murray, J. B., Brown, G. C., Ferrucci, F. & McGuire, W. J. Mechanisms of magma eruption and emplacement at Mt. Etna between 1989 and 1992. *Nature* **361**, 439–441 (1993).
22. Fernandez, J., Tiampo, K. F., Jentzsch, G., Charco, M. & Rundle, J. B. Inflation or deflation? New results for Mayon volcano applying elastic-gravitational modeling. *Geophys. Res. Lett.* **28**, 2349–2352 (2001).
23. Watanabe, H., Okubo, S., Sakashita, S. & Maekawa, T. Drain-back process of basaltic magma in the

- summit conduit detected by microgravity observation at Izu-Oshima volcano, Japan. *Geophys. Res. Lett.* **25**, 2865–2868 (1998).
24. Thouret, J.-C. *et al.* Geomorphological and geological survey, and SPOT remote sensing of the current activity of Nevado Sabancaya stratovolcano (south Peru): assessment for hazard-zone mapping. *Z. Geomorph.* **39**, 515–535 (1995).
25. Wooster, M. J. & Rothery, D. A. Thermal monitoring of Lascar Volcano, Chile, using infrared data from the along-track scanning radiometer: a 1992–1995 time series. *Bull. Volcanol.* **58**, 566–579 (1997).

Acknowledgements

European Space Agency (ESA) remote-sensing satellite (ERS) synthetic aperture radar (SAR) imagery for this study was acquired as a Category 1 research project from the ESA. We thank A. Linde for a critical review, L. Rivera and R. Lohman for modelling software, S. de Silva for an electronic version of his volcano database, and H. Zebker, Y. Fialko, P. Segall, E. Brodsky and M. Battaglia for useful discussions. This material is based upon work partially supported by the National Science Foundation under a grant to M.S. M.E.P. was partly supported by NASA and NSF fellowships.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.E.P. (e-mail: matt@gps.caltech.edu).

Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors

Benjamin Kerr*, Margaret A. Riley†, Marcus W. Feldman* & Brendan J. M. Bohannan*

* Department of Biological Sciences, Stanford University, Stanford, California 94305, USA

† Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06511, USA

One of the central aims of ecology is to identify mechanisms that maintain biodiversity^{1,2}. Numerous theoretical models have shown that competing species can coexist if ecological processes such as dispersal, movement, and interaction occur over small spatial scales^{3–10}. In particular, this may be the case for non-transitive communities, that is, those without strict competitive hierarchies^{3,6,8,11}. The classic non-transitive system involves a community of three competing species satisfying a relationship similar to the children's game rock–paper–scissors, where rock crushes scissors, scissors cuts paper, and paper covers rock. Such relationships have been demonstrated in several natural systems^{12–14}. Some models predict that local interaction and dispersal are sufficient to ensure coexistence of all three species in such a community, whereas diversity is lost when ecological processes occur over larger scales^{6,8}. Here, we test these predictions empirically using a non-transitive model community containing three populations of *Escherichia coli*. We find that diversity is rapidly lost in our experimental community when dispersal and interaction occur over relatively large spatial scales, whereas all populations coexist when ecological processes are localized.

Microbial laboratory communities have proved useful for studying the generation and maintenance of biodiversity^{15–17}. In particular, communities containing toxin-producing (or colicinogenic) *E. coli* have been the centre of much attention from both theoretical ecologists^{3,6,8,18–20} and microbiologists^{21–27}. Colicinogenic bacteria possess a 'col' plasmid, containing genes that encode the colicin (the

toxin), a colicin-specific immunity protein (which renders the cell immune to the colicin) and a lysis protein (which is expressed when the cell is under stress, causing partial cell lysis and the subsequent release of the colicin)^{26,27}. In general, only a small fraction of a population of colicinogenic cells will lyse and release the colicin²⁷. Colicin-sensitive bacteria are killed by the colicin but may occasionally experience mutations that render them resistant to the colicin. The most common mutations alter cell membrane proteins that bind or translocate the colicin^{23,24,26,27}. In some cases, the growth rate of resistant cells (R) will exceed that of colicinogenic cells (C), but will be less than the growth rate of sensitive cells (S). This occurs because resistant cells avoid the competitive cost of carrying the col plasmid^{21,22,26,27} but suffer because colicin receptor and translocation proteins are also involved in crucial cell functions such as nutrient uptake^{21,23,24,26,27}. In such cases, S can displace R (because S has a growth-rate advantage), R can displace C (because R has a growth-rate advantage) and C can displace S (because C kills S). That is, the C–S–R community satisfies a rock–paper–scissors relationship.

Using a modification of the lattice-based simulation of Durrett and Levin⁶, we theoretically explored the role of the spatial scale of

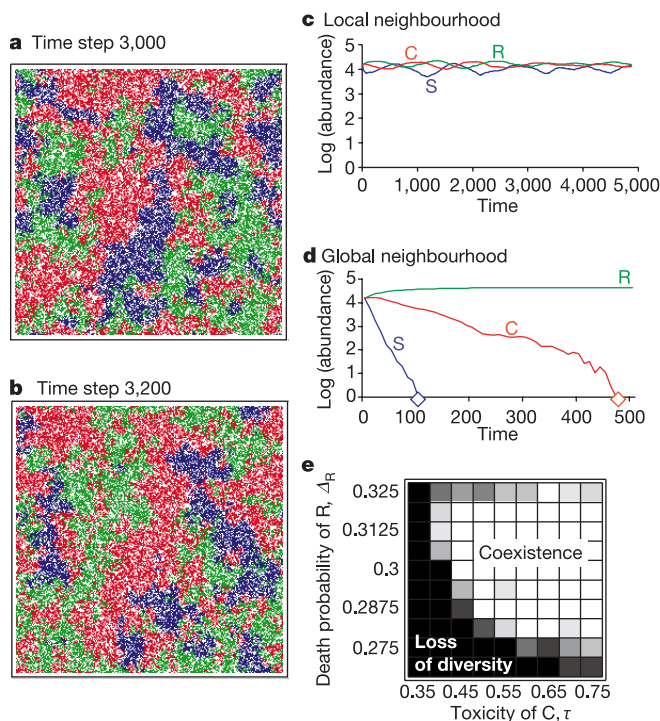


Figure 1 Predictions of the lattice-based simulation (see Box 1). **a, b**, Snapshots of the lattice in a simulation with a local neighbourhood at times 3,000 (**a**) and 3,200 (**b**). The unit of time is an 'epoch', equal to 62,500 lattice point updates (an epoch is the average turnover of any given lattice point in the 250×250 grid). The strains are colour-coded as follows: C is red, S is blue and R is green; empty lattice points are white. **c**, The complete community dynamics for the same simulation run. **d**, Community dynamics for a simulation with a global neighbourhood. The abundances in **c** and **d** are log transformed. When the abundance of a strain goes to zero, we represent this event with a diamond on the abscissa of the relevant graph at the relevant time. For **a–d** we used the following parameters: $\Delta_C = 1/3$, $\Delta_{S,0} = 1/4$, $\Delta_R = 10/32$, and $\tau = 3/4$ (see Box 1). **e**, Sensitivity of qualitative dynamics to changes in a subset of parameter values. For the (τ, Δ_R) values plotted, the greyscale indicates the number of 'local' simulated runs (out of 10) in which coexistence occurred for at least 10,000 epochs, with the lighter area indicating a higher probability of coexistence. For all simulations, we require $\Delta_{S,0} < \Delta_R < \Delta_C < \frac{\Delta_{S,0} + \tau}{1 + \tau}$, which (at least for the mixed system) ensures that S displaces R, R displaces C, and (if C has sufficient density) C displaces S.

ecological processes in our C–S–R community (see Box 1). When dispersal and interaction were local, we observed that ‘clumps’ of types formed (Fig. 1a). These patches chased one another over the lattice—C patches encroached on S patches, S patches displaced R patches and R patches invaded C patches (Fig. 1a, b). Within this fluid mosaic of patches, the local gains made by any one type were soon enjoyed by another type. The result of this balanced chase was the maintenance of diversity (Fig. 1c). However, this balance was lost when dispersal and interaction were no longer exclusively local (that is, in the ‘well-mixed’ system—see Box 1). In the mixed system, continual redistribution of C rapidly drove S extinct, and then R outcompeted C (Fig. 1d). Durrett and Levin⁶ describe a qualitatively similar effect of spatial scale in their model of colicinogenic, sensitive, and ‘cheater’ strains (where a cheater was defined as a strain producing less colicin at a lower competitive cost).

When ecological processes were local in the simulation, coexistence occurred over a substantial range of model parameter values (Fig. 1e), suggesting that the result was not very sensitive to the specific choice of parameter values. In the case of the mixed system, coexistence never occurred for the region of parameter space shown in Fig. 1e. In agreement with Durrett and Levin⁶, our simulation results suggested that three strains with the abovementioned non-hierarchical relationship could coexist when dispersal and interaction are local, whereas one strain excludes the others when the community is well mixed.

To test this conclusion, we used three strains of the bacterium *E. coli*: a colicin-producing strain (C), a sensitive strain (S), and a

resistant strain (R), which satisfied a rock–paper–scissors competitive relationship (see Methods). We placed the C–S–R community in the following three environments: (1) ‘Flask’ (a well-mixed environment in which dispersal and interaction are not exclusively local); (2) ‘Static Plate’ (an environment in which dispersal and interaction are primarily local); and (3) ‘Mixed Plate’ (an environment intermediate between these two extremes).

For the Flask environment, the bacteria were grown in shaken flasks containing liquid media. We transferred an aliquot of the community to fresh media every 24 h. In the Static Plate environment, the bacteria were grown on the surface of solid media in Petri plates. Every 24 h, we pressed each plate onto a platform covered with a sterile velveteen cloth and then placed a fresh plate on the velvet. This method transferred a small sample of the community and allowed the transferred sample to retain the spatial pattern that developed on the previous plate. The Mixed Plate environment was identical to the Static Plate environment, except that at each transfer the fully-grown community plate was pressed on the velvet several times, each time rotated at a different angle (see Methods).

Figure 2a shows that C, S and R strains were maintained at high densities in the Static Plate environment throughout the experiment. Photographs of the plates show the spatial pattern that developed over the experiment (Fig. 3a). The pink and yellow inter-strain boundaries in Fig. 3b show clearly that R chased C, and C

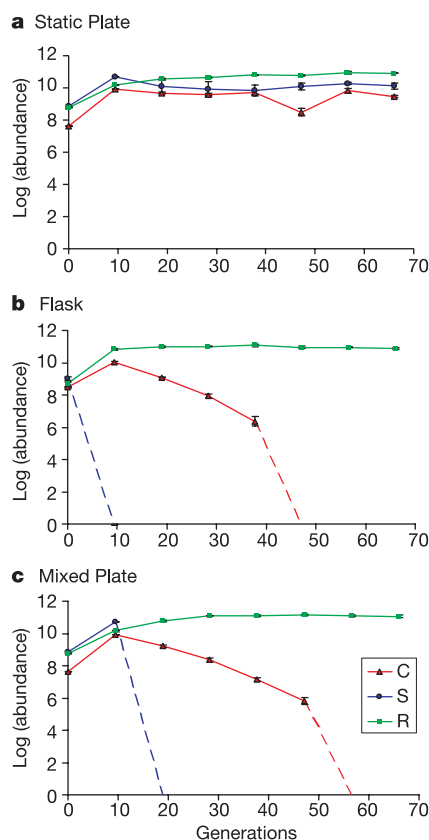


Figure 2 Community dynamics in the experimental treatments: **a**, Static Plate; **b**, Flask; and **c**, Mixed Plate. Dashed lines indicate that the abundance of the relevant strain has decreased below its detection limit. Data points are the mean of three replicates, and bars depict standard errors of the mean. Consecutive data points are separated by 24 h, approximately 10 bacterial generations.

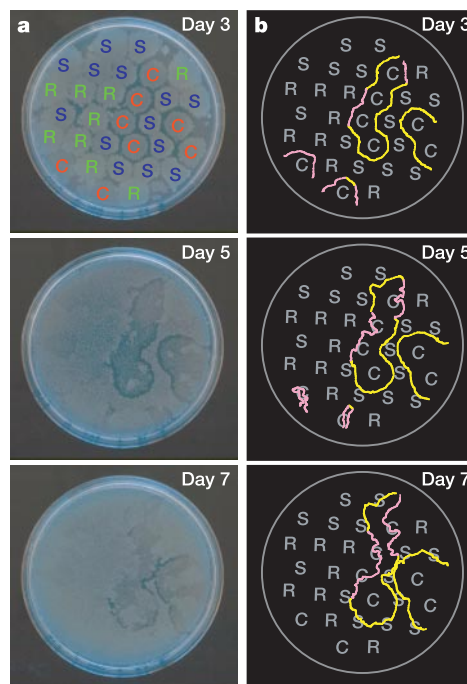


Figure 3 Time series photographs of a representative run of the Static Plate environment. We initiated the plate environments by depositing small droplets from pure cultures in a hexagonal lattice pattern, where the strain at each point was assigned at random. **a**, The changing spatial configuration of the experimental community is shown in this first panel of photographs. Patches inhabited by C cells were less dense and consequently easily distinguished from S and R patches. The dense growing ‘spots’ that appear inside the C clumps were determined to be resistant cells generated *de novo* from S cells. An empty layer existed between C clumps and S clumps, where diffused colicin had prevented the growth of S cells, but where C cells had not yet colonized. The border between C and R lacked this empty layer. **b**, ‘Chasing’ between clumps is highlighted in this second panel. The letters giving the initial spatial distribution of the strains are preserved for reference. The borders between C and S are coloured in yellow and the borders between C and R in pink.

chased S, respectively. It was not possible to see S chasing R, because these two strains grew to comparable densities. However, the S density did not consistently decrease over time (Fig. 2a), suggesting that S did indeed chase R. Thus, coexistence did not result from absolute spatial isolation of the three strains. As can be seen in Fig. 3, interaction (for example, competition and killing) occurred at the boundaries between the strains. These results support the prediction that balanced chasing in a spatially structured, non-hierarchical community may result in the maintenance of diversity.

The results for the Flask environment are shown in Fig. 2b. In contrast to the Static Plate environment, both S and C dropped below their detection limits before the study was completed. The difference in dynamics between the Flask and Static Plate environments did not merely reflect differences between liquid culture and surface growing conditions. Simply mixing the surface growing community at each transfer event (that is, our Mixed Plate environment) produced dynamics very similar to those in the Flask environment (Fig. 2c). Thus, it appears that dispersal and interactions must be local for coexistence to occur in this community.

To explore the robustness of our results, we repeated the experimental work described above using different initial spatial configurations and starting strain frequencies (see Methods). In the cases considered, we observed dynamics very similar to those observed in our original experiment—that is, chasing of types, coexistence of all three types for at least 66 generations when ecological processes were localized, and displacement of S and C and persistence by R when processes were not localized (data not shown). These observations suggested that our results were not due to the particular starting conditions chosen in our initial experiments.

The work described here lies at the crossroads of two recent trends in the study of biodiversity. The first is increasing interest in the role of the spatial scale of ecological processes^{2,4,9,17}. It has been shown theoretically that simply allowing interaction and dispersal to occur

locally can promote diversity within a community^{2,4–9,11}. Our experimental results confirm these predictions. The second trend concerns the role of non-hierarchical competitive relationships^{3,6,11,28,29}. Recent theory suggests that the number of species coexisting on a limited number of resources can be astonishingly high when there are non-hierarchical relationships among the species^{28,29}. Our study system represents a very simple non-transitive triplet: a game of rock–paper–scissors. Our experimental results are consistent with the prediction that non-hierarchical competitive relations can promote diversity. However, in our case, the localization of ecological processes is also necessary.

Localized processes may be important for coexistence in other natural communities as well. For many relatively sessile organisms such as plants, some marine invertebrates and the microbial constituents of biofilms, dispersal and interaction tend to occur over small spatial scales. Communities with such organisms may also exhibit non-transitive relationships. For instance, if one species produces a toxin (a widespread phenomenon, occurring in many species of plants³⁰, marine invertebrates¹³, fungi³⁰, and essentially every major bacterial lineage²⁷) there will be the potential for non-transitivity. This can occur if toxin production is costly, both sensitive and resistant species exist, and costs associated with resistance are less than those associated with toxin production. However, the relative magnitude of these costs will be critical to the prospects for coexistence. Specifically, if the costs of resistance or allelopathy are either too great or too small, the community can collapse, even if the members maintain a non-transitive relationship (for example, see the dark portions of Fig. 1e). If the relative costs in our experimental community are representative of those in natural communities, then we might expect the scale of ecological processes to determine the likelihood of biodiversity maintenance in vegetation with allelopathic plants, coral reefs with toxic sessile invertebrates, and microbial communities with antibiotic-producing microbes. Such communities may be ideal systems to study further the role of local dispersal and interaction in species coexistence. □

Methods

The experimental system

We used the *E. coli* colicin E2 system: strains BZB1011 (S), E2^C-BZB1011 (C), and E2^R-BZB1011 (R). C was marked with resistance to T6 coliphage and S with resistance to T5 coliphage. Other than the non-conjugative col plasmid and markers, C was isogenic to S. R bacteria were generated by selecting marked S mutants able to grow in the presence of the colicin²⁴. All types were counted by selective plating, with S counted by subtraction. Periodically, we checked the markers by verifying that T6-resistant bacteria produced colicin E2 and that T5-resistant bacteria did not.

Determining strain relationships

The relative fitnesses of the strains were determined by growing two competing strains in pairwise competition, both in 10 ml Luria–Bertani (LB) liquid culture and in 4 ml of soft agar on the surface of agar plates. The fitness (w) of strain x relative to that of strain y is

$$w(x, y) = \frac{\ln(x_F/x_0)}{\ln(y_F/y_0)}$$

where x_0 and y_0 are initial densities and x_F and y_F are densities after 24 h of growth.

We found $w(R, S) = 0.59 \pm 0.097$ and $w(R, C) = 1.91 \pm 0.399$ in the liquid culture environment and $w(R, S) = 0.73 \pm 0.142$ and $w(R, C) = 1.90 \pm 0.230$ on the plates. We did not compute $w(S, C)$ in either environment, but we observed that the colicin produced by strain C effectively kills S cells (for example, spotting colicin on a S lawn gives large clear plaques). Thus, C kills S, S outgrows R, and R outgrows C.

Tracking community dynamics

In the Static Plate environment, the bacteria grew on the surface of 4 ml of soft LB agar, which had been poured onto a plate with a hard LB agar base. To initialize a plate, 15 μ l droplets of pure culture were placed in a 31-point hexagonal lattice (Fig. 3a). The droplet pattern was generated by randomly assigning the identity of the strain at each lattice point according to the probability distribution: $\Pr\{C\} = 0.25$, $\Pr\{S\} = 0.5$, and $\Pr\{R\} = 0.25$. After 24 h of growth at 37 °C, we pressed a fully grown plate on a platform covered with velvet. Then three fresh plates were pressed on the velvet, taking care to preserve orientation. The first plate was used for counting strain densities after the next day's growth. This was done by scraping off the soft agar layer with the bacterial community into

Box 1

Lattice-based simulation

We embedded our virtual C–S–R community in a 250×250 regular square lattice with periodic (or ‘wrap-around’) boundaries. To start the simulations shown in Fig. 1, every lattice point was randomly and independently assigned one of the following states: occupation by a C, S or R cell or the ‘empty state’. We used an asynchronous updating scheme, which consisted of sequentially picking random focal points and probabilistically changing their states. The state transition probabilities of a focal point depend both on its current state and the states of points within its neighbourhood. The size of a neighbourhood represents the spatial scale of ecological processes. To explore the role of spatial scale, we simulated our community using two different neighbourhood sizes. The ‘local’ neighbourhood consisted of the eight lattice points directly surrounding a focal point; in such a case, interaction (that is, killing and competition for space) and dispersal (that is, ‘birth’) are local. The ‘global’ neighbourhood consisted of every point in the grid outside of a focal point; in this case, interaction and dispersal are no longer exclusively local, and the community behaves like a well-mixed system. For either neighbourhood, if an empty lattice point is selected for updating, the probability that it is filled with a cell of type i (with $i \in \{C, S, R\}$) is given by f_i , the fraction of its neighbourhood occupied by strain i . If an occupied lattice point in state i is selected, it is killed with probability Δ_i . Although Δ_C and Δ_R are fixed values, Δ_S is not; it is equal to $\Delta_{S,0} + \tau(f_C)$, where $\Delta_{S,0}$ is the probability of death of an S cell without any neighbouring C cells, and τ measures the toxicity of neighbouring C cells.

10 ml of saline, vortexing the mixture for several seconds, and then following standard selective plating protocol. The second plate was used for propagating the community (that is, for pressing on the velvet after the next day's growth). The third plate was photographed (see Fig. 3).

The Mixed Plate environment was identical to the Static Plate environment described above with the following exception: at each transfer, the fully-grown plate was (1) pressed lightly on the velvet, (2) turned clockwise at a randomly chosen angle and pressed a second time, (3) turned randomly counter-clockwise and pressed a third time, and (4) turned randomly clockwise and pressed a fourth time. The fresh plates were then pressed on the velvet to initiate this 'mixed up' sample.

Our Flask environment was a 125 ml flask with 33.75 ml of LB broth shaken at 125 rev min⁻¹ at 37 °C. After 24 h of growth, 50 µl of the culture was transferred to a flask with fresh media. These volumes guaranteed that the total number of bacterial cells in the flask after a full day's growth, and total number of cells transferred, matched the corresponding numbers of cells from the plate runs.

In all three environments, the densities of each type were determined every 24 h, and all treatments were replicated in triplicate. We terminated the experiments after one week (approximately 66 generations), because by this time we detected a substantial number of resistant mutants derived from the S population (for example, the dense clumps within the C patches in Fig. 3a). Evolution of R from S violates an assumption of our model and can lead to a change in the predicted dynamics, especially if the cost of resistance is much lower in the *de novo* R mutants (see low values of Δ_R in Fig. 1e). We are currently exploring such evolutionary phenomena both theoretically and experimentally. It should be noted that one week suffices for a single 15-µl droplet placed in the centre of a plate to nearly cover the plate under the Static Plate regime.

Additional starting conditions

We manipulated the starting strain frequencies and initial spatial configuration on the plates in two additional sets of runs (data not shown). First, we repeated the experiments as outlined above with the same initial hexagonal lattice pattern, but using another random assignment of the strains (taken from the probability distribution above). Second, rather than initiating the plates with droplets in a hexagonal pattern, we mixed the bacterial strains in soft agar at different starting frequencies (with S in excess and C and R rare) and poured the agar over the plates, resulting in an initially 'unclumped' distribution.

Received 12 December 2001; accepted 27 March 2002; doi:10.1038/nature00823.

- Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31**, 343–366 (2000).
- Tilman, D. & Pacala, S. in *Species Diversity in Ecological Communities* (eds Ricklefs, R. E. & Schluter, D.) 13–25 (Univ. Chicago Press, Chicago, 1993).
- Czárán, T. L., Hoekstra, R. F. & Pagie, L. Chemical warfare between microbes promotes biodiversity. *Proc. Natl Acad. Sci. USA* **99**, 786–790 (2002).
- Dieckmann, U., Law, R. & Metz, J. A. J. (eds) *The Geometry of Ecological Interactions: Simplifying Spatial Complexity* (Cambridge Univ. Press, Cambridge, 2000).
- Durrett, R. & Levin, S. The importance of being discrete (and spatial). *Theor. Pop. Biol.* **46**, 363–394 (1994).
- Durrett, R. & Levin, S. Allelopathy in spatially distributed populations. *J. Theor. Biol.* **185**, 165–171 (1997).
- Hassell, M. P., Comins, H. N. & May, R. M. Species coexistence and self-organizing spatial dynamics. *Nature* **370**, 290–292 (1994).
- Pagie, L. & Hogeweg, P. Colicin diversity: A result of eco-evolutionary dynamics. *J. Theor. Biol.* **196**, 251–261 (1999).
- Tilman, D. & Kareiva, P. (eds) *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions* (Princeton Univ. Press, Princeton, 1997).
- Rohani, P., Lewis, T. J., Grunbaum, D. & Ruxton, G. D. Spatial self-organization in ecology: pretty patterns or robust reality? *Trends Ecol. Evol.* **12**, 70–74 (1997).
- Frean, M. & Abraham, E. R. Rock-scissors-paper and the survival of the weakest. *Proc. R. Soc. Biol. Sci.* **B 268**, 1323–1327 (2001).
- Sinervo, B. & Lively, C. M. The rock-paper-scissors game and the evolution of alternative male strategies. *Nature* **380**, 240–243 (1996).
- Buss, L. W. & Jackson, J. B. C. Competitive networks: Nontransitive competitive relationships in cryptic coral reef environments. *Am. Nat.* **113**, 223–234 (1979).
- Paquin, C. E. & Adams, J. Relative fitness can decrease in evolving asexual populations of *S. cerevisiae*. *Nature* **306**, 368–371 (1983).
- Bohannan, B. J. M. & Lenski, R. E. Linking genetic change to community evolution: Insights from studies of bacteria and bacteriophage. *Ecol. Lett.* **3**, 362–377 (2000).
- Korona, R., Nakatsu, C. H., Forney, L. J. & Lenski, R. E. Evidence for multiple adaptive peaks from populations of bacteria evolving in a structured habitat. *Proc. Natl Acad. Sci. USA* **91**, 9037–9041 (1994).
- Rainey, P. B. & Travisano, M. Adaptive radiation in a heterogeneous environment. *Nature* **394**, 69–72 (1998).
- Frank, S. A. Spatial polymorphism of bacteriocins and other allelopathic traits. *Evol. Ecol.* **8**, 369–386 (1994).
- Iwasa, Y., Nakamaru, M. & Levin, S. A. Allelopathy of bacteria in a lattice population: Competition between colicin-sensitive and colicin-producing strains. *Evol. Ecol.* **12**, 785–802 (1998).
- Nakamaru, M. & Iwasa, Y. Competition by allelopathy proceeds in traveling waves: Colicin-immune strain aids colicin-sensitive strain. *Theor. Pop. Biol.* **57**, 131–144 (2000).
- Adams, J., Kinney, T., Thompson, S., Rubin, L. & Helling, R. B. Frequency-dependent selection for plasmid-containing cells of *Escherichia coli*. *Genetics* **91**, 627–637 (1979).
- Chao, L. & Levin, B. R. Structured habitats and the evolution of anticompetitor toxins in bacteria. *Proc. Natl Acad. Sci. USA* **78**, 6324–6328 (1981).

- Feldgarden, M. & Riley, M. A. High levels of colicin resistance in *Escherichia coli*. *Evolution* **52**, 1270–1276 (1998).
- Feldgarden, M. & Riley, M. A. The phenotypic and fitness effects of colicin resistance in *Escherichia coli* K-12. *Evolution* **53**, 1019–1027 (1999).
- Gordon, D. M. & Riley, M. A. A theoretical and empirical investigation of the invasion dynamics of colicinogeny. *Microbiology* **145**, 655–661 (1999).
- James, R., Kleinhous, C. & Moore, G. R. The biology of E colicins: Paradigms and paradoxes. *Microbiology* **142**, 1569–1580 (1996).
- Riley, M. A. & Gordon, D. M. The ecological role of bacteriocins in bacterial competition. *Trends Microbiol.* **7**, 129–133 (1999).
- Huisman, J. & Weissing, F. J. Biodiversity of plankton by species oscillations and chaos. *Nature* **402**, 407–410 (1999).
- Huisman, J., Johansson, A. M., Folmer, E. O. & Weissing, F. J. Towards a solution of the plankton paradox: The importance of physiology and life history. *Ecol. Lett.* **4**, 408–411 (2001).
- Rice, E. L. *Allelopathy* (Academic Press, Orlando, 1984).

Acknowledgements

We thank M. Munos for help in the laboratory, N.B. Raju for helping with the plate photography, and D. Ackerly, P. Armsworth, C. Boggs, C. Devine, P. Godfrey-Smith, D. Gordon, A. Hirsh, J. Huisman, C. Jessup, S. Levin, D. Petrov, P. Rainey, T. Ricketts, S. Tuljapourkar, K. Walag and V. Walbot for many comments on previous versions of the manuscript.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.K. (e-mail: bkerr@stanford.edu).

Exceptional sperm cooperation in the wood mouse

Harry Moore*, Kateřina Dvořáková†, Nicholas Jenkins* & William Breed‡

* Section of Reproductive and Developmental Medicine, University of Sheffield, S10 2UH, UK

† Department of Developmental Biology, Charles University, Prague 2, 128 44, Czech Republic

‡ Department of Anatomical Sciences, University of Adelaide, South Australia SA5005, Australia

Spermatozoa from a single male will compete for fertilization of ova with spermatozoa from another male when present in the female reproductive tract at the same time¹. Close genetic relatedness predisposes individuals towards altruism, and as haploid germ cells of an ejaculate will have genotypic similarity of 50%, it is predicted that spermatozoa may display cooperation and altruism to gain an advantage when inter-male sperm competition is intense². We report here the probable altruistic behaviour of spermatozoa in an eutherian mammal. Spermatozoa of the common wood mouse, *Apodemus sylvaticus*, displayed a unique morphological transformation resulting in cooperation in distinctive aggregations or 'trains' of hundreds or thousands of cells, which significantly increased sperm progressive motility. Eventual dispersal of sperm trains was associated with most of the spermatozoa undergoing a premature acrosome reaction. Cells undergoing an acrosome reaction in aggregations remote from the egg are altruistic in that they help sperm transport to the egg but compromise their own fertilizing ability.

Several examples of sperm cooperation have been reported mainly in molluscs and insects^{3,4}. A possible exception in Mammalia is the spermatozoa of opossums that conjugate to form pairs during sperm maturation and disengage immediately before fertilization⁵. Sperm will benefit from cooperation if Hamilton's rule⁶ is fulfilled. This depends on the probability of sperm survival in terms of

reaching the site of fertilization and the difference in relatedness of cooperating sperm and other sperm competing for fertilization. For true altruism, the fertilizing capacity of one spermatozoon is compromised or sacrificed to benefit another; however, evidence in Eutheria has been largely lacking. Spermatozoa of some rodents (for example, guinea-pig) stack in rouleaux formation⁷ or agglutinate, but these cell associations do not appear particularly advantageous. Conversely, it has been suggested that a primary function of some spermatozoa in the rat and human ejaculate is to incapacitate spermatozoa of another male, so-called kamikaze spermatozoa^{8,9}; however, this hypothesis is not supported by experimental evidence¹⁰.

Among small mammals, multiple matings resulting in sperm competition and mixed paternity in littermates are believed to be widespread¹. The wood mouse, *A. sylvaticus*, is a common murid rodent throughout Western Europe, with a breeding season from February to October¹¹. Adult males were caught in woodland near Sheffield between February and September, or obtained from a breeding colony in captivity (Department of Veterinary Science, University of Liverpool). The weight of wild-caught mice (21.4 ± 2.9 g, $n = 10$) was not significantly different to that of males bred in captivity (22.6 ± 2.1 g, $n = 8$). Mean testes weight of 1.02 ± 0.16 g gave a relative testes/body mass of $4.76 \pm 0.38\%$,

higher than for almost all rodent species and in accordance with high testis sperm output¹². The cauda epididymidis lies underneath a hairless protrusion of the scrotum, a morphological adaptation that may have evolved for efficient cooling and storage of spermatozoa, thereby maximizing ejaculate sperm output¹³. These features are consistent with a high degree of inter-male sperm competition^{1,14}, as also indicated by a radio-tracking study that concluded that male wood mice engage in scramble competition to mate polygynously with promiscuous females¹⁵. This suggests that sperm competition and multiple paternity is probably commonplace in wood mice; however, as far as we are aware, direct DNA parentage studies have not been undertaken.

The murid rodent sperm head usually displays a falciform morphology¹⁶. In the wood mouse, this morphology developed after meiosis and was complete at spermiogenesis (Fig. 1a). Epididymal spermatozoa exhibited an extremely long apical hook composed of an extended perforatorium region¹⁷ (Fig. 1b, c). Uniquely, the hook was attached invariably (>95%) to a peri-nuclear process, spur and neck region on the lower ventral surface of the sperm head by electron-dense adhesive material. Propidium iodide staining indicated that only the basal region of the apical hook was of nuclear origin (Fig. 1c, d). The acrosome was situated over the convex and upper lateral region of the sperm head but extended for only part of

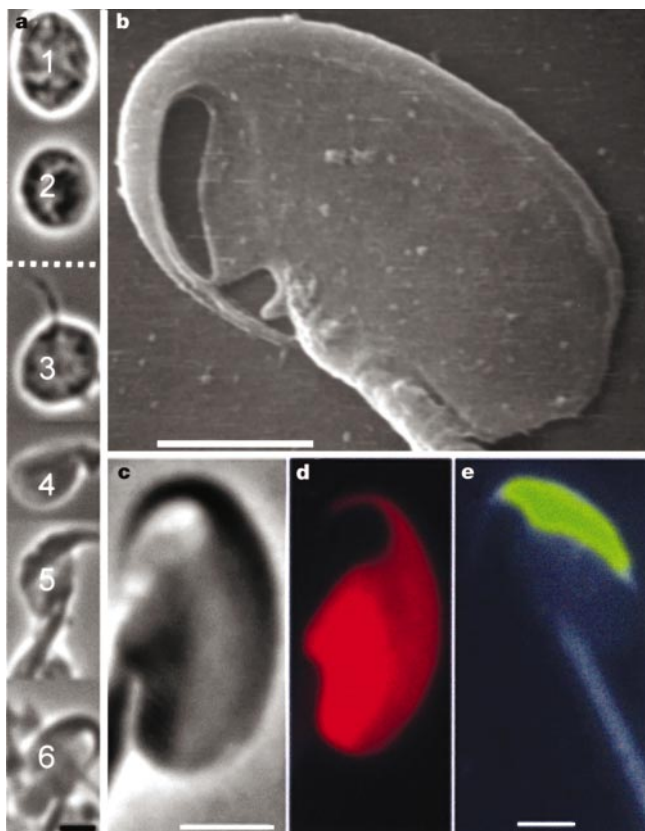


Figure 1 Morphology of wood mouse sperm head. **a**, Montage of germ cells extracted from wood mouse testis. 1, spermatogonia; 2, pachytene spermatocyte. Below the dotted line indicates spermiogenesis; 3, round spermatid with flagellum; 4 and 5, elongating spermatid; 6, testicular spermatozoa at spermiogenesis from epithelium, note completion of apical hook. **b**, Scanning electron micrograph of a sperm head showing attachment of apical hook to ventral shelf, stud and neck region. **c**, Phase-contrast image of sperm head. **d**, Propidium iodide nuclear staining of **c** showing the limit of nuclear contribution to the hook. **e**, Immunofluorescent staining of acrosomal region. Scale bars: **a**, 2 μ m; **b**, 2.5 μ m; **c-e**, 3 μ m.

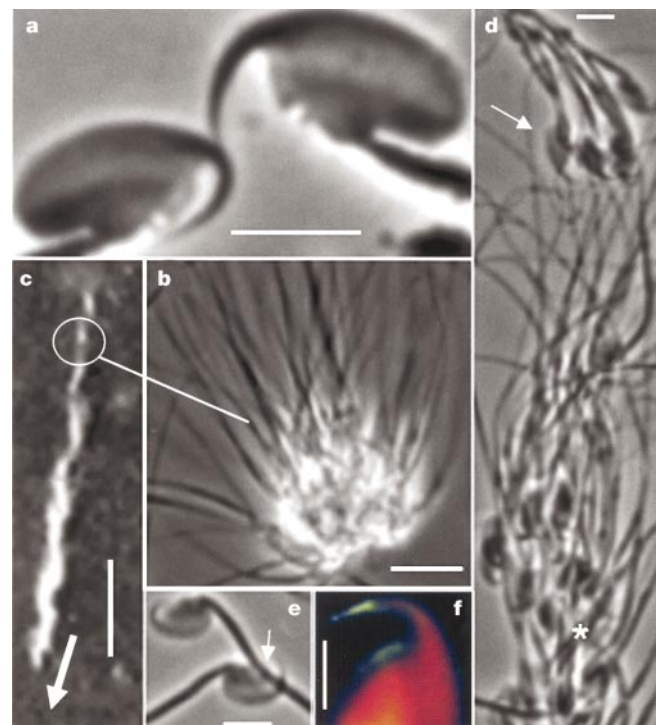


Figure 2 Development of sperm train. **a**, Phase-contrast image of spermatozoa after 2 min in IVF medium. The spermatozoon on the right has fully deployed its hook; compare with the spermatozoon on the left where the hook is still attached laterally to the head. **b**, Motile aggregation of approximately 50 spermatozoa. **c**, Dark-field frame from a video clip of a large motile sperm train over 2-mm in length and consisting of thousands of spermatozoa. In each smaller clump in the line (circled), spermatozoa were aggregated as in **b**. The train moved in the direction of the arrow with rapid sinusoidal motility. **d**, Phase-contrast image of formalin-fixed spermatozoa from a train. Sperm attach by means of apical hook to flagellum (arrow), or apical hook to apical hook (asterisk). **e**, Phase-contrast of a spermatozoon attached by the apical hook to the flagellum of another cell (indicated by an arrow). **f**, Fluorescent (fluorescein isothiocyanate) immunolocalization of filamentous actin in the sperm head with propidium iodide counterstain. Actin was expressed in the apical hook region more intensely after deployment. Scale bars: **a**, **d**, **e**, **f**, 5 μ m; **b**, 10 μ m; **c**, 1 mm.

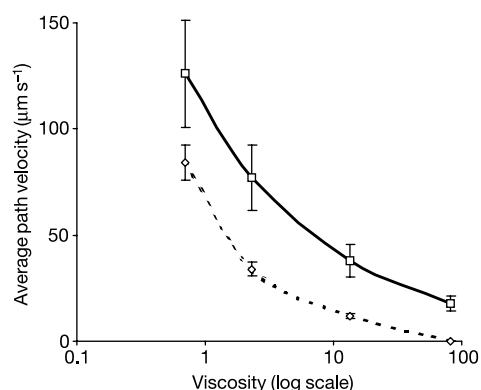


Figure 3 The change in progressive motility (average path velocity) of sperm trains (solid line) and single spermatozoa (dotted line) with increasing viscosity of medium. Sperm trains displayed significantly greater average path velocity ($P \leq 0.01$) than single spermatozoa at all values.

the hook, as visualized with acrosome-specific monoclonal antibody 18.6 (ref. 18) (Fig. 1e).

Cauda epididymal spermatozoa displayed good motility ($\geq 90\%$ progressively motile) when released into laboratory mouse *in vitro* fertilization medium¹⁹ and initially were in single cell suspension. Within 1–5 min, the apical hook of each spermatozoon moved away from the ventral process and spur (Fig. 2a) and intertwined and adhered to other deployed apical hooks or to the flagellum of other spermatozoa, forming motile clumps of 10–50 cells (Fig. 2b). By 5 min most of these clumps ($\geq 85\%$) formed motile ‘trains’ of spermatozoa comprising hundreds to several thousand cells (Fig. 2c, d). These trains exhibited rapid progressive motility, often with a sinusoidal motion (see Supplementary Information). As determined by computer-assisted sperm analysis²⁰, the mean average path velocity of sperm trains ($132 \pm 21 \mu\text{m s}^{-1}$, $n = 200$ per sample) was significantly greater than for single spermatozoa ($87 \pm 24 \mu\text{m s}^{-1}$, $P \leq 0.05$). Initiation of sperm aggregations was concomitant with adhesion of the inner concave surface of the apical hook with flagellum (Fig. 2e) or apical hooks of other spermatozoa. The intensity of staining for filamentous actin (but not other cytoskeletal proteins; data not shown) in the apical hook increased after its deployment, suggesting that this protein might have a function in hook remodelling (Fig. 2f). Cell–cell adhesion between spermatozoa occurred along the inner surface of the apical hook. The difference in progressive motility between single spermatozoa and trains was relatively more pronounced when the viscosity of medium was increased with polyvinyl pyrrolidone (PVP-10) to that estimated for secretions of the female reproductive tract⁵ (Fig. 3). At the highest viscosity (10% PVP-10), progressive motility was exhibited only by sperm trains. Several reports indicate that the mean progressive motility of spermatozoa, as measured *in vitro*, is positively correlated with the outcome of fertilization *in vivo*^{20–22}.

To eliminate the possibility that sperm trains resulted from an artefact of *in vitro* culture conditions, sperm behaviour in the female tract was investigated after mating. Captive-bred adults were maintained in breeding groups of typically four females and two males (14 h light/10 h dark cycle, dusk at 16:00). After mating (0.5–2 h), females ($n = 5$) were killed and their reproductive tract removed and flushed with phosphate buffered saline by fine pipette under a dissecting microscope. Motile sperm trains (approximately 50–200 sperm) and many non-motile single spermatozoa were recovered from the vagina of each female. In three females, motile sperm trains (10–50 sperm) were detected in the uterine lumen but very few ($< 5\%$) single spermatozoa.

Dispersal of the ‘train’ of spermatozoa occurred after about

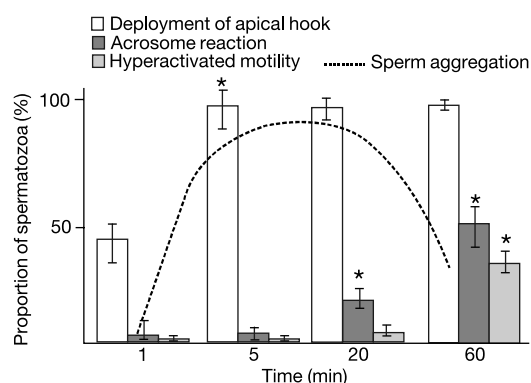


Figure 4 Histogram of the proportion of spermatozoa displaying deployment of the apical hook, the acrosome reaction and hyperactivated motility in relation to sperm aggregation *in vitro* at different time points. Error bars indicate s.e.m. Asterisk, significantly different from previous time point ($P \leq 0.005$; $n = 6$).

30 min and was complete by 90 min in capacitating medium. Spermatozoa retained motility ($\geq 90\%$) but there was a significant increase in the proportion of spermatozoa that had undergone premature loss of the acrosome (as detected by monoclonal antibody 18.6) and showed hyperactivated motility. A time course for these changes in relation to deployment of the apical hook and sperm aggregation is shown in Fig. 4.

Such close cooperation of many spermatozoa to significantly enhance sperm motility is rare and thus far has not been observed in this manner in any mammalian species. We hypothesize that inter-male sperm competition in conjunction with unique morphological features of the sperm head were conducive for evolution of this elaborate gamete behaviour. In the male excurrent ducts, where spermatozoa are densely packed, an exposed elongated apical hook is liable to cause sperm aggregation and blockage, leading to infertility. An adaptation where the hook is initially attached to the lower ventral surface of the sperm head would prevent premature sperm aggregation but subsequent deployment of the hook in the female reproductive tract would promote beneficial sperm aggregation and sperm motility. A ‘green beard’ gene might evoke such cellular cooperation although other genetic mechanisms cannot be excluded at this stage. A green beard is a gene that recognizes copies of itself in other individuals and directs benefits to those individuals²³. Such genes (or groups of linked genes) are predicted to be rare because they must code for a combination of features including a recognizable morphological or behavioural phenotype, a recognition system and a response (in this case enhanced sperm motility particularly in viscous secretion)²⁴. During mammalian spermatogenesis, germ cells proliferate and differentiate while remaining connected by inter-cytoplasmic bridges, allowing gene products to diffuse freely between individual cells²⁵. In such a syncytium, most products of post-meiotic gene expression will tend to be equilibrated between spermatozoa, particularly membrane surface components. Thus, it may be significant that the mechanism in *A. sylvaticus* entails a non-diffusible component, namely the perforatorium, and therefore a modification to a trait could remain specific to the cell. The perforatorium and apical hook clearly developed and attached to the lower ventral sperm head during post-meiotic testis development in the wood mouse. Although some genes involved in the formation of these organelles may be first expressed at the diploid stage, many others such as testis-specific gamma actin have only been detected in post-meiotic germ cells^{26,27}. Single mutations in spermiogenic-specific genes can profoundly affect sperm-head morphology²⁸.

A mechanism for dispersal of aggregated sperm before fertiliza-

tion is required. This was concomitant with induction of the acrosome reaction in a proportion of the aggregated spermatozoa *in vitro*, and acrosomal proteolytic enzymes may be responsible for digestion of cell–cell adhesion molecules. Acrosome-reacted murine spermatozoa lose their capacity to bind to the zona pellucida and fertilize²⁹, and therefore these cells displayed altruistic behaviour. In contrast, acrosome-intact spermatozoa free of aggregation would retain fertilizing capacity²⁹. □

Received 1 March; accepted 23 April 2002; doi:10.1038/nature00832.

1. Birkhead, T. R. & Moller, A. P. *Sperm Competition and Sexual Selection* (Academic, London, 1996).
2. Trivers, R. *Social Evolution* (Benjamin Cummings, California, 1985).
3. Siviniski, J. in *Sperm Competition and the Evolution of Animal Mating Systems* (ed. Smith, R. L.) 223–249 (Academic, Orlando, 1984).
4. Hayashi, F. Insemination through an externally attached spermatophore: bundled sperm and post-copulatory mate guarding by male fishflies (Megaloptera: corydalidae). *J. Insect Physiol.* **42**, 859–866 (1996).
5. Moore, H. D. Gamete biology of the new world marsupial, the grey short-tailed opossum, *Monodelphis domestica*. *Reprod. Fertil. Dev.* **8**, 605–615 (1996).
6. Hamilton, W. D. The genetical evolution of social behaviour. I. *J. Theor. Biol.* **7**, 1–16 (1964).
7. Shepherd, B. A. & Martan, J. Morphology and fertility of guinea-pig spermatozoa aged *in vivo*. *Arch. Androl.* **2**, 53–58 (1979).
8. Baker, R. R. & Bellis, M. A. Kamikaze sperm in mammals? *Anim. Behav.* **36**, 936–939 (1988).
9. Baker, R. R. & Bellis, M. A. *Human Sperm Competition* (Chapman Hall, London, 1995).
10. Moore, H. D., Martin, M. & Birkhead, T. R. No evidence for killer sperm or other selective interactions between human spermatozoa in ejaculates of different males *in vitro*. *Proc. R. Soc. Lond. B* **266**, 2243–2350 (1999).
11. Hanak, V. & Mazak, V. *Mammals* (Aventinum, Prague, 1979).
12. Breed, W. G. & Taylor, J. Body mass, testes mass and sperm size in murine rodents. *J. Mamm.* **81**, 758–768 (2000).
13. Bedford, J. M. in *The Spermatozoon* (eds Fawcett, D. W. & Bedford, J. M.) 7–21 (Urban and Schwarzenberg, Baltimore, 1979).
14. Harcourt, A. H., Harver, P. H., Larson, S. G. & Short, R. V. Testes weight, body weight and breeding systems in primates. *Nature* **293**, 55–57 (1981).
15. Tew, T. E. & Macdonald, D. W. Dynamics of space use and male vigor amongst wood mice, *Apodemus sylvaticus* in the cereal ecosystem. *Behav. Ecol. Sociobiol.* **34**, 337–345 (1994).
16. Yanagimachi, R. in *Physiology of Reproduction* (eds Knobil, E. & Neil, J.) 189–317 (Raven Press, New York, 1994).
17. Friend, G. F. The sperms of British muridae. *Quart. J. Micro. Sci.* **78**, 419–443 (1936).
18. Moore, H. D., Hartman, T. D., Brown, A. C., Smith, C. A. & Ellis, D. A. Expression of sperm antigens during spermatogenesis and maturation detected with monoclonal antibodies. *Exp. Clin. Immunogenet.* **2**, 84–96 (1985).
19. Fraser, L. R. & Drury, L. M. The relationship between sperm concentration and fertilisation *in vitro* of mouse eggs. *Biol. Reprod.* **13**, 513–518 (1975).
20. Moore, H. D. & Akhondji, M. A. Fertilizing capacity of rat spermatozoa is correlated with decline in straight-line velocity measured by continuous computer-aided sperm analysis. *J. Androl.* **17**, 50–60 (1996).
21. Holt, C., Holt, W. V., Moore, H. D., Reed, H. C. & Curnock, R. M. Objectively measured boar sperm parameters correlate with the outcome of on-farm inseminations. Results of two fertility trials. *J. Androl.* **18**, 312–323 (1996).
22. Birkhead, T. R., Martinez, J. G., Burke, T. & Froman, D. P. Sperm mobility determines the outcome of sperm competition in the domestic fowl. *Proc. R. Soc. Lond. B* **266**, 1–6 (1999).
23. Dawkins, R. *The Selfish Gene* (Oxford Univ. Press, Oxford, 1976).
24. Haig, D. Gestational drive and green-bearded placenta. *Proc. Natl Acad. Sci. USA* **93**, 6547–6551 (1996).
25. De Kretser, D. M. & Kerr, J. B. in *Physiology of Reproduction* (eds Knobil, E. & Neil, J.) 1177–1290 (Raven Press, New York, 1994).
26. Kleene, K. C. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mech. Dev.* **106**, 3–23 (2001).
27. Kim, E., Waters, S. H., Hake, L. E. & Hecht, N. B. Identification and developmental expression of a smooth muscle gamma-actin in post-meiotic male germ cells of mice. *Mol. Cell Biol.* **9**, 1875–1881 (1989).
28. Xu, X., Toselli, P. A., Russell, L. D. & Seldin, D. C. Globozoospermia in mice lacking the casein kinase II α catalytic subunit. *Nature Genet.* **23**, 118–121 (1999).
29. Wassarman, P. M., Jovine, L. & Litscher, E. S. A profile of fertilization in mammals. *Nature Cell Biol.* **3**, E59–E64 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank J. Waters for advice on wood mice husbandry. K.D. was supported by an Erasmus Exchange studentship.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.M. (e-mail: h.d.moore@shef.ac.uk).

Transmitter-evoked local calcium release stabilizes developing dendrites

Christian Lohmann, Karen L. Myhr & Rachel O. L. Wong

Department of Anatomy and Neurobiology, Washington University School of Medicine, 660 S. Euclid, St Louis, Missouri 63110, USA

In the central nervous system, dendritic arborizations of neurons undergo dynamic structural remodelling during development. Processes are elaborated, maintained or eliminated to attain the adult pattern of synaptic connections^{1–3}. Although neuronal activity influences this remodelling^{4–6}, it is not known how activity exerts its effects. Here we show that neurotransmission-evoked calcium (Ca^{2+}) release from intracellular stores stabilizes dendrites during the period of synapse formation. Using a ballistic labelling method to load cells with Ca^{2+} indicator dyes⁷, we simultaneously monitored dendritic activity and structure in the intact retina. Two distinct patterns of spontaneous Ca^{2+} increases occurred in developing retinal ganglion cells—global increases throughout the arborization, and local ‘flashes’ of activity restricted to small dendritic segments. Blockade of local, but not global, activity caused rapid retraction of dendrites. This retraction was prevented locally by focal uncaging of caged Ca^{2+} that triggered Ca^{2+} release from internal stores. Thus, local Ca^{2+} release is a mechanism by which afferent activity can selectively and differentially regulate dendritic structure across the developing arborization.

To observe directly how neural activity dynamically regulates dendritic structure during development, we introduced Ca^{2+} indicator dyes into retinal ganglion cells by a ballistic method that resulted in complete filling of the dendritic arborization⁷ (Fig. 1a). Using Oregon Green 488 BAPTA-1 dextran (OGB1; $k_d = 0.3$ – $0.4 \mu\text{M}$), Ca^{2+} levels were monitored across the dendritic arborization of embryonic days (E) 8–19 chick retinal ganglion cells ($n = 93$ cells), during the period of dendritic remodelling and synaptogenesis. Two distinct types of Ca^{2+} transients were observed (Fig. 1b and c; see Supplementary Information 1). First, local patches of dendrites showed transient increases in intracellular Ca^{2+} concentration ($[\text{Ca}^{2+}]_i$) that extended from 5.0 to 16.0 μm (mean $\pm$ s.e.m. = $9.6 \pm 0.9 \mu\text{m}$; 5 cells, 12 events) and lasted between 3.0 and 28.0 s (8.9 ± 2.5 s; 7 cells, 26 events). The mean change in fluorescence ($\Delta F/F_0$; see Methods) of these events was $27.0 \pm 1.4\%$ (12 cells, 129 events). With a lower affinity indicator, Fluo-4 dextran ($k_d = 4.1 \mu\text{M}$, $n = 3$ cells, 5 events; see Supplementary Information 2A), we observed local events with similar duration (10.2 ± 1.8 s) and extent ($9.8 \pm 2.9 \mu\text{m}$), albeit at lower amplitude ($\Delta F/F_0 = 13.2 \pm 1.9\%$). Local flashes appeared with a frequency of $5.2 \pm 0.7 \text{ min}^{-1}$ per mm dendrite at E13. Qualitative comparison of the onset times of local events across the arborization did not reveal any strict temporal patterns in their occurrence, although they can occur repeatedly in the same locations (Fig. 1e).

In addition to local flashes, we observed Ca^{2+} levels simultaneously increasing in the soma and throughout the arborization (Fig. 1b and c). Global activity was periodic at intervals of 80.0 ± 16.0 s, occurring synchronously in neighbouring cells. In contrast to local flashes, global activity was shorter in duration (2.0 – 7.0 s, mean 3.9 ± 0.7 s, 4 cells, 10 events) and relatively larger in amplitude ($\Delta F/F_0$: $53.0 \pm 4.8\%$, 12 cells, 23 events). The global activity resembled that of spontaneous bursting activity observed during propagation of retinal waves^{8–10}. Simultaneous calcium imaging and perforated patch recordings (5 cells) demonstrated

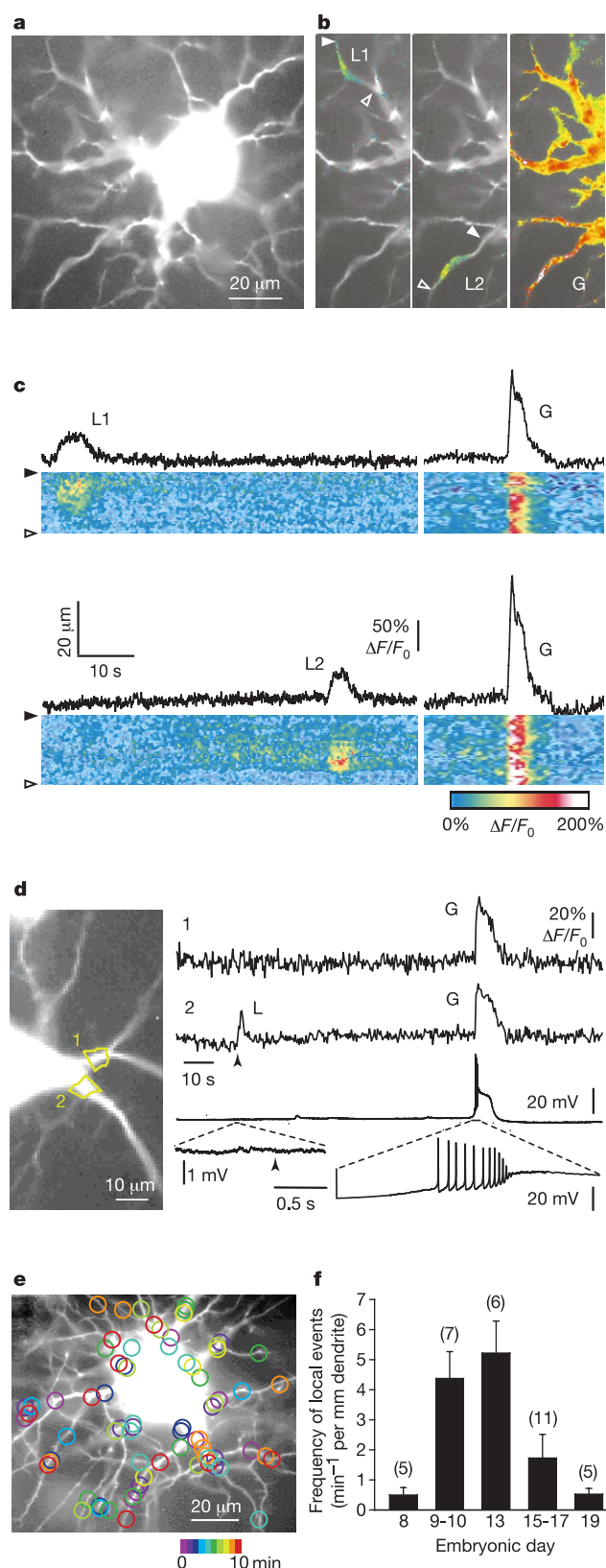


Figure 1 Developing dendrites of retinal ganglion cells exhibit global and local spontaneous $[\text{Ca}^{2+}]_i$ increase. **a**, E13 cell filled with OGB1 dextran by ballistic delivery. **b**, Local (L1, L2) and global (G) Ca^{2+} increases at three times. Increases in $[\text{Ca}^{2+}]_i$ represented by $\Delta F/F_0$ (pseudocolour scale in **c**). **c**, Relative timing of the global and the two local events in **b**. Arrowheads mark extent of the measured dendritic segments. **d**, Simultaneous Ca^{2+} imaging and perforated-patch voltage recording (E13). Arrowhead, onset of local event. **e**, Distribution of local events of this cell. **f**, Local activity with age; error bars show $\pm$ s.e.m; number of cells shown in parentheses.

that global events were associated with depolarization and bursts of action potentials (24 events; Fig. 1d). Voltage changes were not detected during local events (15 events; Fig. 1d). All ganglion cells aged E13–19, and some before E13, exhibited global activity. Local flashes, however, occurred most frequently at E9–13 (Fig. 1f), during the period of extensive dendritic growth and remodelling⁵ that coincides with synaptogenesis in the inner retina^{11,12}.

The mechanisms underlying local and global activities were distinct. We focused on E13, when local activity was at its peak. Ca^{2+} influx was necessary for both local and global activities. However, local but not global activity was significantly reduced by thapsigargin (Tha) and cyclopiazonic acid (CPA) which inhibit the endoplasmic reticulum Ca^{2+} ATPases (Table 1). Local activity was also diminished by 2-aminoethoxydiphenyl borate (2-APB) and ryanodine (Rya) which block InsP_3 - and ryanodine-sensitive stores, respectively (Table 1). Neurotransmission contributes to evoking this release (Table 1). The neuronal nicotinic receptor antagonist dihydro- β -erythroidine hydrobromide (DH β E) reduced local but not global activity. Atropine, a non-selective antagonist of muscarinic receptors, modulated the frequency of both activities. Local and global activities were unaffected by blocking ionotropic glutamate receptors with 1,2,3,4-tetrahydro-6-nitro-2,3-dioxo-benzo[f]quinoxaline-7-sulphonamide (NBQX) and 2-amino-5-phosphonovaleic acid (APV), or with the metabotropic glutamate receptor antagonist, methyl-(4-carboxyphenyl) glycine (MCPG). Although acetylcholine-mediated and not glutamate-mediated transmission evokes Ca^{2+} release at E13, it remains possible that the latter transmission plays a part at other ages. Local activity persisted in the presence of tetrodotoxin (TTX), which blocked global activity^{8,10}. The persistence of local activity in TTX is not surprising because bipolar cell and most amacrine cell transmission are independent of action potentials. Together, these results suggest that at E13, nicotinic receptor-mediated transmission contributes significantly to evoking local release of Ca^{2+} from internal stores in the ganglion cell dendrites.

Because acetylcholine-mediated transmission arises from amacrine cells, it is possible that the local flashes occur near sites of amacrine cell contact. In one case where we fortuitously labelled a displaced amacrine cell and a nearby ganglion cell (Fig. 2a and b), we observed a potential contact form (Fig. 2c, and Supplementary Information 3). Upon initial contact, the dendrite of the ganglion cell did not show any change in $[\text{Ca}^{2+}]_i$ at the contact site. However, an hour later, the dendritic region adjacent to the site of contact exhibited spontaneous local increases in $[\text{Ca}^{2+}]_i$ (Fig. 2d and e). Although we do not know if acetylcholine-mediated transmission was involved, ionotropic glutamate-receptor-mediated transmission was not involved because the recording was performed in the presence of NBQX and APV. We thus suspect that local flashes are associated with sites at which stable contacts have formed with

Table 1 Distinct mechanisms trigger local and global activity

Treatment	Local events		Global events		n
	Frequency (%)	Amplitude (%)	Frequency (%)	Amplitude (%)	
Zero calcium	$-96 \pm 2^{**}$	-	$-100 \pm 0^{**}$	-	7
Thapsigargin (1 μ M)	$-84 \pm 6^{**}$	$-41 \pm 11^*$	3 ± 13	-1 ± 7	6
CPA (20 μ M)	$-88 \pm 4^*$	-	-15 ± 10	-	5
Ryanodine (10 μ M)	$-76 \pm 8^{**}$	-	12 ± 23	-	5
2-APB (100 μ M)	$-76 \pm 11^*$	-	-15 ± 41	-	4
DH β E (100 μ M)	$-63 \pm 9^{**}$	45 ± 33	-11 ± 15	36 ± 17	6
Atropine (3 μ M)	$-27 \pm 8^*$	-	$-29 \pm 11^*$	-	6
NBQX (10 μ M) + APV (100 μ M)	-6 ± 15	-	9 ± 22	-	8
MCPG (1 mM)	-27 ± 12	-	7 ± 19	-	6
TTX (1 μ M)	-15 ± 15	-12 ± 8	$-85 \pm 9^{**}$	-11 ± 33	5

See text for definitions of abbreviations in column 1. * $P < 0.05$; ** $P < 0.01$; n, number of cells.

amacrine cells, possibly acetylcholine-containing amacrine cells, which constitute 90% of displaced amacrine cells in the chick retina¹³.

Because local signalling occurs during the period of dendritic development, we examined whether disruption of local activity altered dendritic organization. Dendrites retracted rapidly upon application of thapsigargin or 2-APB (Fig. 3a and b). Images of cells ($n = 2$) acquired every minute suggested that retraction started within a few minutes of thapsigargin application (Supplementary Information 4). Similar retraction occurred in the presence of all pharmacological agents that blocked local activity and Ca^{2+} release from stores (Fig. 3c, Supplementary Information 2B). The retraction rate decreased with time, perhaps because the remaining larger proximal processes are more stable than the finer terminal dendrites. In contrast, dendrites were stable in the presence of pharmacological agents that preserved local activity, even when global activity was blocked (TTX, Fig. 3c). Indeed, *in vivo* blockade with TTX does not affect dendritic development of retinal ganglion cells¹⁴. Together, these observations indicate that local, not global, activity sustains dendrites during development.

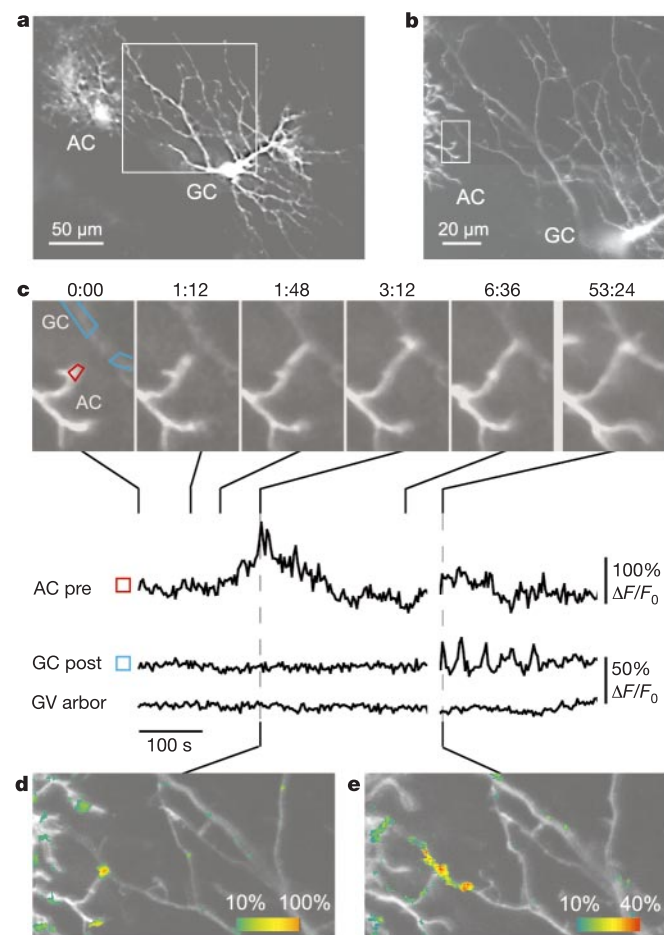


Figure 2 Potential contact between a displaced amacrine cell (AC) neurite and a ganglion cell (GC) dendrite. **a**, Confocal reconstructions of the two cells labelled with OGB1 dextran. **b**, Higher magnification of area of process overlap. **c**, Images at times (min:s) corresponding to the $\Delta F/F_0$ plots for: AC pre, tip of the amacrine process (red region; upon contact, $\Delta F/F_0$ was adjusted for background fluorescence of the underlying dendrite); GC post, $\Delta F/F_0$ averaged across the blue regions; GC arbor, $\Delta F/F_0$ averaged across all dendrites of the ganglion cell shown in **b** outside GC post. **d**, **e** $\Delta F/F_0$ upon contact (**d**) and 52 min after contact (**e**).

To address directly whether dendrites can be selectively stabilized by local Ca^{2+} release, we attempted to differentially 'rescue' dendrites from retracting when spontaneous local flashes were blocked. Caged Ca^{2+} (*o*-nitrophenyl EGTA, NP-EGTA) was introduced together with OGB1 into single cells using the ballistic method. In this way, Ca^{2+} release was restricted to the labelled cell. Co-labelling with the Ca^{2+} indicator allowed us to estimate the spread of Ca^{2+} that was uncaged within the dendritic arborization. Before uncaging, the retinae were placed in zero Ca^{2+} solution (Ringer's solution containing 0 mM Ca^{2+} , 4 mM Mg^{2+} and 1 mM BAPTA) to induce net retraction of dendrites. Ca^{2+} was then released locally by focusing an ultraviolet (UV) laser beam upon part of the arborization (Fig. 4a). Uncaging caused a local but relatively sustained elevation in $[\text{Ca}^{2+}]_i$ when the stores were not pharmacologically blocked. In contrast, in the presence of ryanodine and 2-APB, laser uncaging evoked a localized but transient increase in $[\text{Ca}^{2+}]_i$ (Fig. 4b). These observations suggest that uncaging in the absence of store blockers may have caused Ca^{2+} -induced Ca^{2+} release (CICR). Terminal dendrites that experienced CICR after photo-uncaging did not retract, whereas unstimulated processes retracted

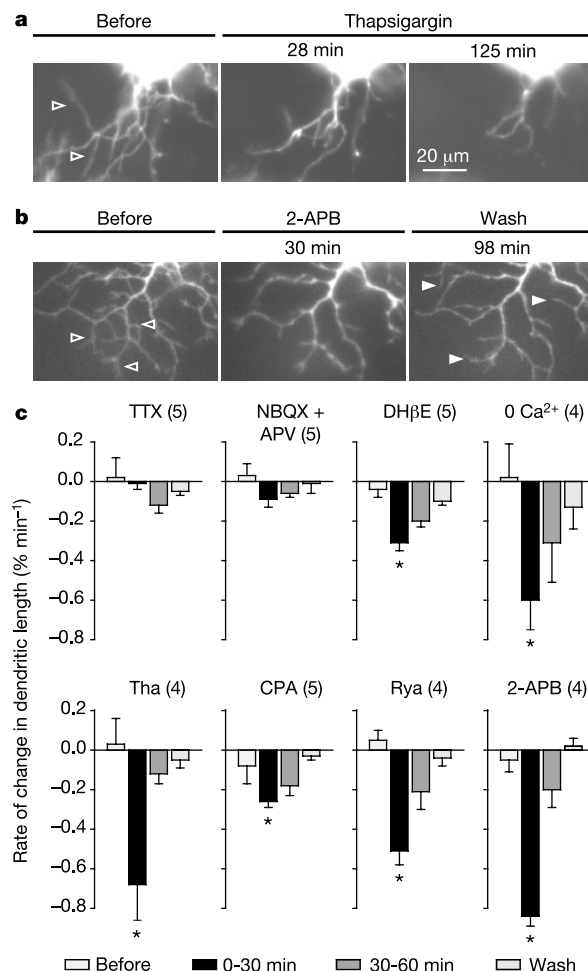


Figure 3 Dendrites retract when local activity is suppressed. **a**, **b**, Examples of dendritic retraction in OGB1-labelled cells (E13) in the presence of thapsigargin or 2-APB. Open arrowheads indicate some terminal dendrites that retracted upon drug application. Filled arrowheads in **b** indicate processes that appeared upon washout. **c**, Summary of the effects of agents with different actions on local and global activities. See text for abbreviations and Table 1 for concentrations of agents; parentheses indicate number of cells. Asterisk, $P < 0.05$, compared to control (before) conditions.

as expected (Fig. 4c and d, Supplementary Information 2c). Continual activity may be required for stabilizing dendrites, because dendrites within the stimulated region retracted somewhat after photo-uncaging ceased (Fig. 4d). When uncaging occurred in the presence of ryanodine and 2-APB, both stimulated and unstimulated dendrites retracted (Fig. 4e). Thus, CICR is required to locally stabilize dendritic processes.

What regulatory signals could local activity provide that differ from those of global activity? Differences in amplitude, temporal and spatial properties of the Ca^{2+} signal can account for the

enormous diversity of actions by this ion^{15,16}. The local and global activities differ in all these characteristics, and in the mechanisms by which $[\text{Ca}^{2+}]_i$ is increased within the dendrite. CICR, rather than Ca^{2+} influx alone, is the key mechanism stabilizing dendrites: the localized nature and/or sustained increase in $[\text{Ca}^{2+}]_i$ of CICR may be the important factors. Despite the large increase in $[\text{Ca}^{2+}]_i$ during a global event, $[\text{Ca}^{2+}]_i$ rises due to influx might be limited to regions close to the cell membrane, and do not effectively unload the stores. Alternatively, global events may trigger Ca^{2+} release, but opening of voltage-gated Ca^{2+} channels concurrent with nicotinic receptor activation may suppress the downstream effects of CICR, as previously observed¹⁷. Local release of Ca^{2+} in cultured cells¹⁸ has been shown to regulate spine structure¹⁹ and synaptic plasticity^{20,21}. The spatial confinement of Ca^{2+} increase in spines²² and small segments of dendrites enables local regulation of synaptic strength^{23,24}. Likewise, localized Ca^{2+} release may dynamically and selectively regulate dendritic structure across the arborization during development by maintaining only dendrites that receive active inputs^{25,26}, thus enabling an efficient distribution of postsynaptic material. Local control of dendritic stability may also be a means for selecting between competing inputs; inputs that trigger CICR may be maintained whereas those that fail to do so may be eliminated by dendritic retraction. The patterning of synaptic inputs onto a dendritic tree may thus be sculpted not only by selective withdrawal or maintenance of afferent terminals, but also by local stabilization or destabilization of dendrites. □

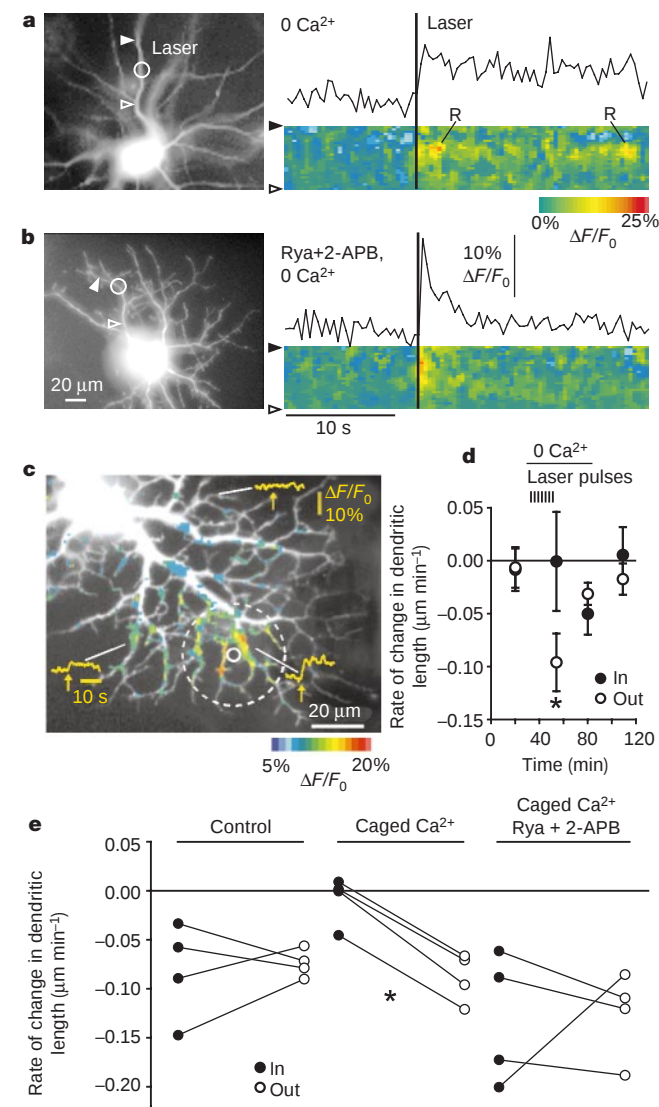


Figure 4 Ca^{2+} -induced Ca^{2+} release prevents dendritic retraction. **a**, **b**, E13 cells filled with OGB1 and NP-EGTA (caged Ca^{2+}), and plots of $\Delta F/F_0$ with time. Circles, centre of laser spot; arrowheads, spatial extent of the measured dendritic segments. Putative Ca^{2+} -induced Ca^{2+} release (R). **c**, Evoked $[\text{Ca}^{2+}]_i$ increase (maximal within dotted circle) given by $\Delta F/F_0$, averaged over 10 s after stimulation. Traces of $\Delta F/F_0$ in three dendrites upon laser stimulation (arrows). **d**, Dendritic measurements of cell in **c** within (in, $n = 12$ processes) and outside (out, $n = 24$) the dotted circle before, during, and after zero Ca^{2+} Ringer's condition. Laser pulses (vertical lines) started 3 min after introducing zero Ca^{2+} solution. Asterisk, $P < 0.01$, comparing unstimulated (out) processes before and during zero Ca^{2+} . **e**, Summary of effects for stimulated versus unstimulated dendrites in zero Ca^{2+} . Connected symbols represent data from one cell. (Asterisk, caged- Ca^{2+} , in versus out: $P < 0.001$).

Methods

Tissue preparation and dye-filling

Embryonic chick retinae were prepared as previously described¹⁰. Before Ca^{2+} imaging, the retina was shot with tungsten particles coated with a Ca^{2+} indicator using the 'Calistics' method⁷. Briefly, 1.5 mg of OGB1 (Molecular Probes, no. O-6798) or Fluo-4 dextran (Molecular Probes, no. F-14240) was dissolved in H_2O and then mixed with 25 mg of 1.7- μm -diameter tungsten particles (BioRad). The mixture was spread thinly on a glass slide, allowed to dry completely, scraped and transferred into polypropylene tubing (BioRad) that was precoated with 1% (w/v) polyvinylpyrrolidone (Sigma) solution. For the uncaging studies, the particles were co-coated with 0.5 mg NP-EGTA (Molecular Probes, no. N-6802) and 1.5 mg OGB1. Particles were delivered at 80 p.s.i. through a culture insert membrane (Falcon, no 35-3092) using the BioRad Helios gene gun²⁷. Cells labelled by Calistics maintained their endogenous content and activity over several hours. Cells were not damaged using the ballistic method because (1) they exhibited spontaneous Ca^{2+} activity resembling retinal wave activity¹⁰, (2) their dendritic filopodia moved at rates similar to that previously observed in cells expressing green fluorescent protein⁵ and (3) synaptic excitatory potentials were observed.

Imaging

Recordings were performed using a cooled CCD (SensiCam, Cooke) or a silicon-intensifier-target tube (SIT) camera (C2400-08, Hamamatsu), controlled with MetaFluor or MetaMorph (Universal Imaging Inc.). Images were acquired at 0.3–3 Hz (CCD) or 7.5–30 Hz (SIT), for continuous recording over 1–10 min (CCD) or 0.5–2 min (SIT). Integration time for each image was 0.2–2.5 s (CCD). To release Ca^{2+} focally, a UV laser beam (Stabilite 2017, Spectra Physics) was focused onto a selected region of the dendritic arborization using the 40 $\times$ water immersion objective. The half-maximal width of the spot was 32–36 μm . The power of the UV beam at the sample was less than 1 mW. $[\text{Ca}^{2+}]_i$ levels were monitored during photo-uncaging of NP-EGTA. A dichroic mirror placed in front of the mercury lamp housing reflected the UV laser beam into the objective, and enabled blue excitation light to pass through during Ca^{2+} imaging. Typically, seven consecutive pulses (10–40 ms duration, approximately 8–32 μJ per pulse) were delivered at 3-min intervals—this was the minimal stimulus that produced reliable uncaging with each pulse. Pharmacological agents were obtained from Sigma Chemicals unless stated otherwise.

Patch recording

Perforated-patch voltage recordings (gramicidin D, 100 $\mu\text{g l}^{-1}$, Sigma) were carried out in current-clamp mode, with no current injected²⁸. Excitatory postsynaptic potentials (EPSPs) were identified using Minianalysis (Synaptosoft)²⁸. EPSPs were defined as events with rise times of < 1 ms (typically 0.3 ms), and amplitudes that exceeded the root mean square of the baseline noise (typically around 0.08 mV) by threefold. No EPSPs were detected within 1 s before or 0.5 s after the onset of the local Ca^{2+} events.

Data analysis

Changes in $[\text{Ca}^{2+}]_i$ are given by $\Delta F/F_0$ where F_0 is the baseline fluorescence level, ΔF is the

change in intensity from this value. Care was taken to omit all changes that were due to focal shifts or sample movement in any dimension. The spatial extent (half-maximal width) of local flashes was measured from line scans of $\Delta F/F_0$ at the time of maximal intensity change. To determine the effects of the pharmacological agents, we followed and compared the summed length of the same processes in control Ringer's solution and after applying the agents. Because the distribution of neuronal processes is largely two-dimensional in the retina, complete representations of the majority of dendrites could be obtained in single planes. Comparison of the total length of the processes in control solution at two time points, 20 min apart, indicated that addition and retraction of processes were balanced, with no net change in total length²⁵. The total length was then measured 25–30 min after the drug was introduced. Changes in total length upon drug treatment were expressed as a percentage of baseline length. The rate of change was calculated by dividing this number by the time between measurements (% change per min). In the uncaging experiments, we measured the total length of all terminal processes in focus that were at least partially located within 20 μ m around the centre of the laser spot, before and after stimulation. All other terminal dendrites outside this region of maximal stimulation served as controls. The average initial lengths of individual terminal dendrites were similar for both groups (inside and outside the stimulated area). All statistical analyses: *t*-test, paired or unpaired.

Received 23 March; accepted 1 May 2002; doi:10.1038/nature00850.

1. Yuste, R. & Tank, D. W. Dendritic integration in mammalian neurons, a century after Cajal. *Neuron* **16**, 701–703 (1996).
2. Cline, H. T. Dendritic arbor development and synaptogenesis. *Curr. Opin. Neurobiol.* **11**, 118–126 (2001).
3. Wong, W. T. & Wong, R. O. L. Rapid dendritic movements during synapse formation and rearrangement. *Curr. Opin. Neurobiol.* **10**, 118–124 (2000).
4. Li, Z., Van Aelst, L. & Cline, H. T. Rho GTPases regulate distinct aspects of dendritic arbor growth in *Xenopus* central neurons *in vivo*. *Nature Neurosci.* **3**, 217–225 (2000).
5. Wong, W. T., Faulkner-Jones, B. E., Sanes, J. R. & Wong, R. O. L. Rapid dendritic remodeling in the developing retina: dependence on neurotransmission and reciprocal regulation by Rac and Rho. *J. Neurosci.* **20**, 5024–5036 (2000).
6. Wong, W. T. & Wong, R. O. L. Changing specificity of neurotransmitter regulation of rapid dendritic remodeling during synaptogenesis. *Nature Neurosci.* **4**, 351–352 (2001).
7. Kettunen, P. *et al.* Rapid loading of calcium indicators by particle mediated ballistic delivery. *J. Neurosci. Methods*, in the press.
8. Catsicas, M., Bonness, V., Becker, D. & Mobbs, P. Spontaneous Ca^{2+} transients and their transmission in the developing chick retina. *Curr. Biol.* **8**, 283–286 (1998).
9. Sernagor, E., Eglen, S. J. & O'Donovan, M. J. Differential effects of acetylcholine and glutamate blockade on the spatiotemporal dynamics of retinal waves. *J. Neurosci.* **20**, RC56 (2000).
10. Wong, W. T., Sanes, J. R. & Wong, R. O. L. Developmentally regulated spontaneous activity in the embryonic chick retina. *J. Neurosci.* **18**, 8839–8852 (1998).
11. Hughes, W. F. & LaVelle, A. On the synaptogenic sequence in the chick retina. *Anat. Rec.* **179**, 297–302 (1974).
12. Hering, H. & Kröger, S. Formation of synaptic specializations in the inner plexiform layer of the developing chick retina. *J. Comp. Neurol.* **375**, 393–405 (1996).
13. Millar, T. J. *et al.* Cholinergic amacrine cells of the chicken retina: a light and electron microscope immunocytochemical study. *Neuroscience* **21**, 725–743 (1987).
14. Wong, R. O. L., Herrmann, K. & Shatz, C. J. Remodeling of retinal ganglion cell dendrites in the absence of action potential activity. *J. Neurobiol.* **22**, 685–697 (1991).
15. Berridge, M. J. Neuronal calcium signalling. *Neuron* **21**, 13–26 (1998).
16. Spitzer, N. C., Lautermilch, N. J., Smith, R. D. & Gomez, T. M. Coding of neuronal differentiation by calcium transients. *Bioessays* **22**, 811–817 (2000).
17. Chang, K. T. & Berg, D. K. Voltage-gated channels block nicotinic regulation of CREB phosphorylation and gene expression in neurons. *Neuron* **32**, 855–865 (2001).
18. Koizumi, S. *et al.* Characterization of elementary Ca^{2+} release signals in NGF-differentiated PC12 cells and hippocampal neurons. *Neuron* **22**, 125–137 (1999).
19. Korkotian, E. & Segal, M. Release of calcium from stores alters the morphology of dendritic spines in cultured hippocampal neurons. *Proc. Natl Acad. Sci. USA* **96**, 12068–12072 (1999).
20. Svoboda, K. & Mainen, Z. F. Synaptic $[Ca^{2+}]_i$: intracellular stores spill their guts. *Neuron* **22**, 427–430 (1999).
21. Rose, C. R. & Konnerth, A. Stores not just for storage. Intracellular calcium release and synaptic plasticity. *Neuron* **31**, 519–522 (2001).
22. Yuste, R., Majewska, A. & Holthoff, K. From form to function: calcium compartmentalization in dendritic spines. *Nature Neurosci.* **3**, 653–659 (2000).
23. Eilers, J., Augustine, G. J. & Konnerth, A. Subthreshold synaptic Ca^{2+} signalling in fine dendrites and spines of cerebellar Purkinje neurons. *Nature* **373**, 155–158 (1995).
24. Nishiyama, M., Hong, K., Mikoshiba, K., Poo, M. M. & Kato, K. Calcium stores regulate the polarity and input specificity of synaptic modification. *Nature* **408**, 584–588 (2000).
25. Vaughn, J. E. Fine structure of synaptogenesis in the vertebrate central nervous system. *Synapse* **3**, 255–285 (1989).
26. McAllister, A. K. Cellular and molecular mechanisms of dendrite growth. *Cereb. Cortex* **10**, 963–973 (2000).
27. Gan, W. B., Grutzendler, J., Wong, W. T., Wong, R. O. L. & Lichtman, J. W. Multicolor “DiOlistic” labeling of the nervous system using lipophilic dye combinations. *Neuron* **27**, 219–225 (2000).
28. Myhr, K. L., Lukasiewicz, P. D. & Wong, R. O. L. Mechanisms underlying developmental changes in the firing patterns of ON and OFF retinal ganglion cells during refinement of their central projections. *J. Neurosci.* **21**, 8664–8671 (2001).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank J. Sanes, J. Lichtman, A. M. Craig and S. Eglen for comments on the manuscript,

and J. Demas for assistance with ballistic loading of calcium indicators. This work was supported by NIH, NSF, DFG and the Plum foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.O.L.W. (e-mail: wongr@pcg.wustl.edu).

TRPV3 is a calcium-permeable temperature-sensitive cation channel

Haoxing Xu^{*†}, I. Scott Ramsey^{*†}, Suhas A. Kotecha^{*†}, Magdalene M. Moran[‡], Jayhong A. Chong^{*}, Deborah Lawson[§], Pei Ge[§], Jeremiah Lilly[§], Inmaculada Silos-Santiago[§], Yu Xie[§], Peter S. DiStefano[§], Rory Curtis[§] & David E. Clapham^{*}

^{*} Howard Hughes Medical Institute, Children's Hospital, Harvard Medical School, Enders 1309, 320 Longwood Avenue, Boston, Massachusetts 02115, USA

[‡] Program in Neuroscience, Harvard Medical School, Goldenson Building B2-228, 220 Longwood Avenue, Boston, Massachusetts 02115, USA

[§] Millennium Pharmaceuticals Inc., 75 Sidney Street, Cambridge, Massachusetts 02139, USA

[†] These authors contributed equally to this work

Transient receptor potential (TRP) proteins are cation-selective channels that function in processes as diverse as sensation and vasoregulation. Mammalian TRP channels that are gated by heat and capsaicin ($>43^{\circ}\text{C}$; TRPV1 (ref. 1)), noxious heat ($>52^{\circ}\text{C}$; TRPV2 (ref. 2)), and cooling ($<22^{\circ}\text{C}$; TRPM8 (refs 3, 4)) have been cloned; however, little is known about the molecular determinants of temperature sensing in the range between $\sim 22^{\circ}\text{C}$ and 40°C . Here we have identified a member of the vanilloid channel family, human TRPV3 (hTRPV3) that is expressed in skin, tongue, dorsal root ganglion, trigeminal ganglion, spinal cord and brain. Increasing temperature from 22°C to 40°C in mammalian cells transfected with hTRPV3 elevated intracellular calcium by activating a nonselective cationic conductance. As in published recordings from sensory neurons, the current was steeply dependent on temperature, sensitized with repeated heating, and displayed a marked hysteresis on heating and cooling^{5–10}. On the basis of these properties, we propose that hTRPV3 is thermosensitive in the physiological range of temperatures between TRPM8 and TRPV1.

Survival depends on the precise regulation of body temperature¹¹. Certain dorsal root ganglia (DRG) neurons detect temperature changes between 32°C and 43°C (warmth)⁵ and respond with calcium influx processes⁷. Recordings of a nonselective ion channel activated by temperatures from 25°C to 49°C (ref. 6), as well as temperature-activated cationic currents in both central and sensory neurons^{8–10} are indicative of an involvement of TRP ion channels. Activation of TRP channels depolarizes cells from the resting membrane potential and shortens action potential duration. Most TRP channels are cation-nonselective and permeant to the signal transduction element Ca^{2+} . Although all channels and enzymes are inherently temperature-sensitive, the discovery of ‘hot’ pepper¹, noxious heat⁸ and menthol-sensitive^{3,4} TRP channels with high 10-degree temperature coefficients (Q_{10} values (ref. 12)) motivates the search for other ion channels in the mammalian TRP superfamily.

|| Present address: Elixir Pharmaceuticals Inc. 1000, One Kendall Square, Cambridge Massachusetts 02139, USA.

Database searches revealed TRPV-related expressed sequence tags (ESTs) in a human testis library. The corresponding genomic sequences were identified on human chromosome 17 near the TRPV1 locus (chromosome location 17p13). A combination of exon-prediction software, polymerase chain reaction (PCR), rapid amplification of cloned ends (RACE) and screening of a human brain complementary DNA library yielded a putative open reading frame (ORF) of 790 amino acids with homology to members of the TRPV family (Fig. 1a). The hTRPV3 cDNA is 38%, 38%, 32% and 12% identical to that of TRPV4, TRPV1, TRPV2 and TRPM8, respectively. Hidden Markov modelling¹³ of the transmembrane topology of the amino acid sequence of hTRPV3 was consistent with a primary structure containing six transmembrane domains (Fig. 1b). As is characteristic for other TRPV and TRPC family channels (Fig. 1c), hTRPV3 contains three putative ankyrin repeats (Fig. 1a).

TaqMan quantitative PCR with reverse transcription (RT-PCR) of 22 different human tissues revealed high TRPV3 expression in brain, spinal cord, DRG, skin and testis. Lower expression was seen in stomach, trachea, small intestine and placenta (see Supplementary Information). Expression of an approximately 5-kilobase (kb) transcript in tongue and stomach was also detected by northern blot analysis (see Supplementary Information). *In situ* hybridization of monkey tissue with an antisense RNA probe derived from the 3' end

of the coding region of TRPV3 labelled neurons throughout the cortex, thalamus and striatum (Fig. 2; see also Supplementary Information). In the spinal cord, TRPV3 was present in ventral horn motoneurons, as well as interneurons in the intermediate zone and deep laminae of the dorsal horn, but not in the superficial dorsal horn (laminae I and II). In the peripheral nervous system, TRPV3 was detected in sympathetic neurons in the superior cervical ganglia (SCG) and in most of the sensory neurons in the DRG and trigeminal ganglia (Fig. 2). Bright field microscopy revealed that TRPV3 RNA is present in DRG sensory neurons of all sizes, including the small nociceptive neurons that previously were shown to express the heat-sensitive channel TRPV1 and the cold-sensor channel TRPM8 (refs 1, 3, 4, 14). No signal was detected in glial cells in either the central or peripheral nervous system. Finally, TRPV3 was detected in the upper layer of monkey tongue (data not shown) and in cells surrounding hair follicles of human skin that receive sensory and sympathetic innervation¹⁵.

Given the homology of hTRPV3 to other heat-activated channels, we tested whether hTRPV3 coded for a functional heat-sensitive and Ca^{2+} -permeable channel. One defining characteristic for temperature-gated channels is a rapid, nonlinear change as a functional response to temperature changes. We transfected hTRPV3 into CHO-K1 (Chinese hamster ovary) cells and measured intracellular calcium ($[\text{Ca}^{2+}]_i$) concentration responses and hTRPV3 currents

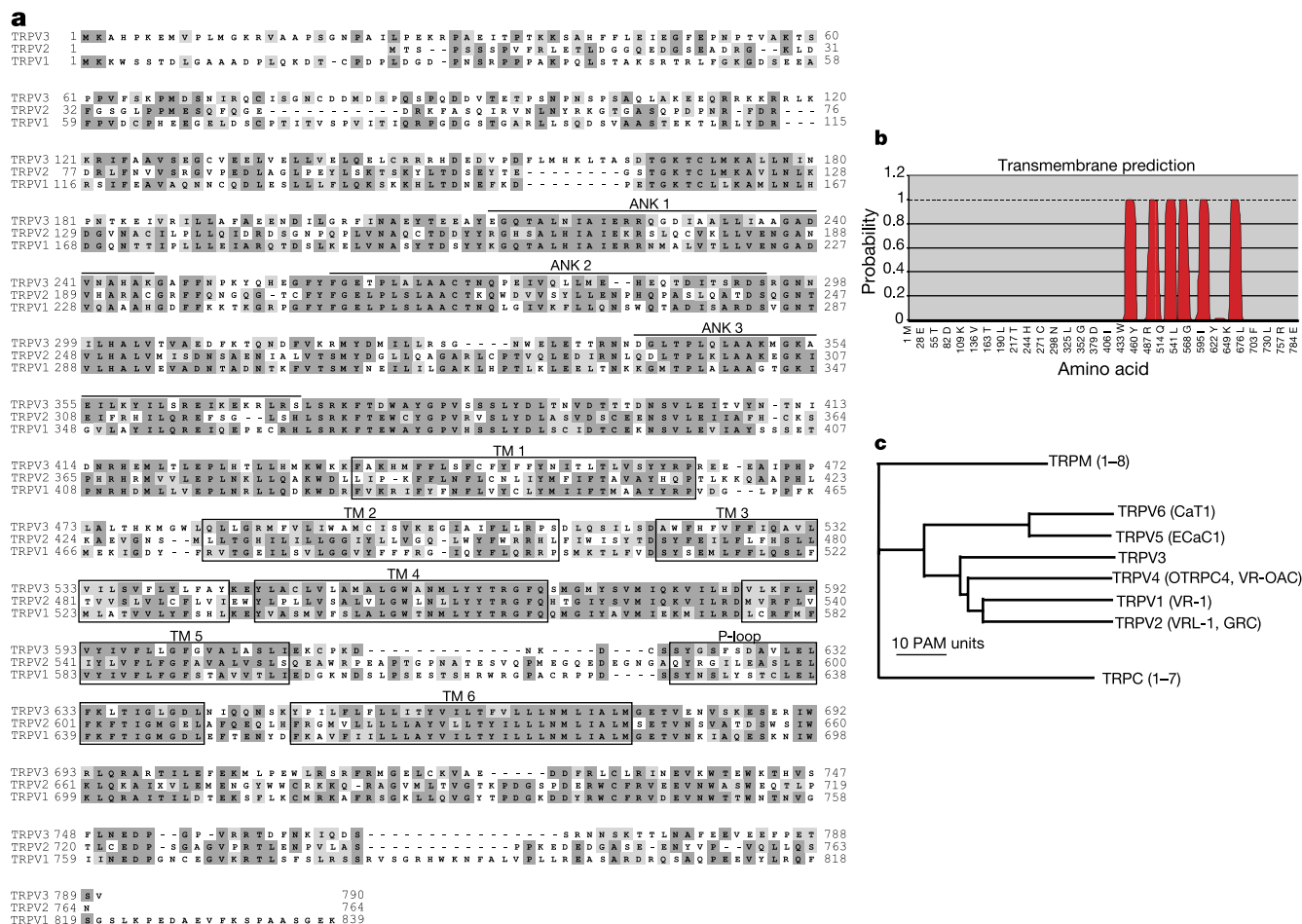


Figure 1 Deduced primary amino acid sequence of hTRPV3 and alignment with hTRPV1 and hTRPV2. **a**, The putative six transmembrane domains (TM 1–6), pore region, and 3 amino-terminal ankyrin repeats are indicated. The carboxy terminus contains a TRP family sequence (IWRLQR) and a consensus PDZ binding motif, ETSV. Identical and homologous residues are shaded in dark and light grey, respectively. GenBank accession numbers for

sequences used in the alignment are NP_542437 (hTRPV1), NP_057197 (hTRPV2) and AF514998 (hTRPV3). **b**, Hidden Markov modelling predicts six transmembrane domains (1–6) and a pore region (P). **c**, Phylogenetic tree of the TRP ion channel superfamily. PAM is defined as 'point accepted mutation' in the MacVector software, Oxford Molecular Group.

(I_{hTRPV3}) 24 h later. Cells were loaded with the Ca^{2+} -sensitive fluorescent dye Fura-2 AM, and fluorescence ratios recorded during perfusion of a heated bath solution. Vector-transfected CHO-K1 cells exhibited no appreciable increase in $[Ca^{2+}]_i$ in response to a temperature ramp. In contrast, $[Ca^{2+}]_i$ in cells transfected with hTRPV3 exhibited an average peak increase of 398 ± 94 nM (mean $\pm$ s.e.m.) in response to a temperature ramp from 22°C to 38°C over 12–15 s ($n = 34$ cells, Fig. 3a, b). The response decayed slowly when the temperature was maintained at a plateau temperature. Reduction of extracellular calcium to nominal concentrations (~ 10 μ M) reduced $[Ca^{2+}]_i$ to basal levels (60 ± 20 nM; Fig. 3a, $n = 26$ cells).

Increasing bath temperature from 23°C to 36°C (initial rate = 0.7°C s^{-1}) elicited a large outwardly rectifying whole-cell current (I_{hTRPV3} , Fig. 3c–f). I_{hTRPV3} readily activated above room temperature ($22.6 \pm 0.1^\circ\text{C}$) and rapidly deactivated after removal

of the heating stimulus (Fig. 3c). The average heat-activated I_{hTRPV3} was 1.4 ± 0.3 nA at +80 mV and -0.42 ± 0.12 nA at -80 mV (mean $\pm$ s.e.m.). We frequently observed that a residual component of I_{hTRPV3} decayed slowly after cooling from peak temperatures (Fig. 3c, d). Similar heating protocols applied to vector- and non-transfected cells failed to activate a similar current (Fig. 3c). I_{hTRPV3} reversed near 0 mV (Fig. 3d), as expected for nonselective cation channels. Large, fast-current fluctuations about the mean I_{hTRPV3} were observed at both negative and positive membrane potentials (Fig. 3c, d), a characteristic of neuronal heat-activated currents⁸.

No currents were measured in cells transfected with hTRPV3 when agonists of TRPV1 (10 μ M capsaicin, 10 μ M resiniferatoxin, pH 4.3) or TRPV2 (10 nM insulin-like growth factor-I)¹⁶ were applied at 22°C. Activators of TRPV4 (hypotonic solution 250 mosM^{17,18} and 10 μ M phorbol ester (PDD¹⁹)) also failed to

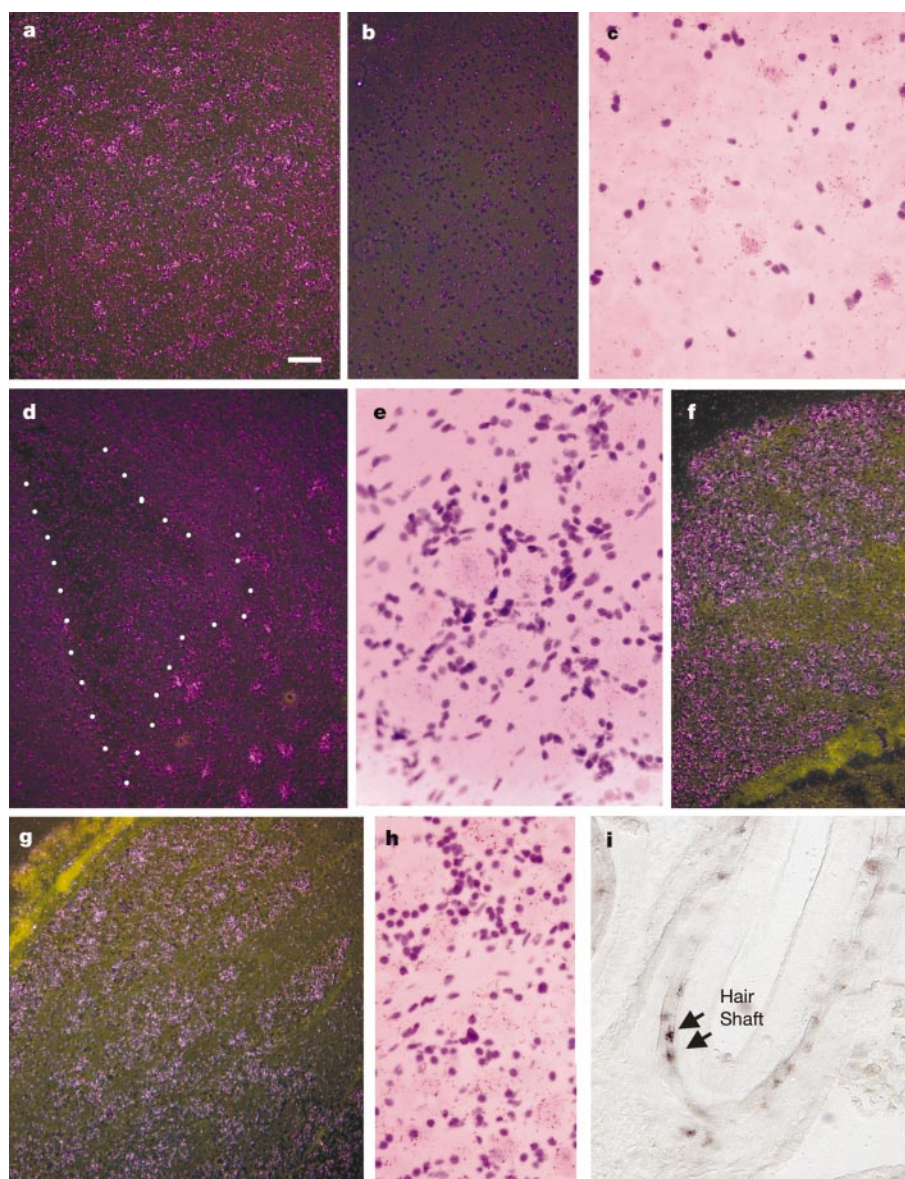


Figure 2 *In situ* hybridization of monkey tissue sections showing expression of TRPV3 in central and peripheral neurons and in the skin. **a–i**, Emulsion autoradiograms of monkey cortex (**a**), cortex sense probe control (**b**), thalamus (**c**), dorsal spinal cord (**d**), trigeminal ganglion (**e**), superior cervical ganglion (**f**), dorsal root ganglion (**g**, **h**), and human hair

follicle root sheath cells (arrows (**i**)). Note the lack of expression in the superficial laminae of the dorsal horn (dotted line in **d**). Scale bar: 40 μ m for $\times 10$ magnification dark field images (**a**, **b**, **d**, **f**, **g**); 10 μ m for $\times 40$ magnification bright field images (**c**, **e**, **h**, **i**).

activate any new currents in hTRPV3-transfected cells at 22 °C. To exclude contamination by native Mg^{2+} -inhibited current^{20–22}, we commonly studied the temperature-dependent activation of I_{hTRPV3} in the presence of 10 mM $[Mg^{2+}]_i$.

I_{hTRPV3} was cation-nonselective (Fig. 3e) with relative permeabilities of $Ca^{2+} > Na^+ \approx K^+ \approx Cs^+$ and permeability ratios as listed in Fig. 3f. Ruthenium red (10 μ M), a nonselective blocker of many cation channels, blocked heat-evoked I_{hTRPV3} by $89 \pm 3\%$ (holding potential, $V_h = -40$ mV, $n = 5$, $P < 0.05$). In contrast, the TRPV1 antagonist capsaizipine (10 μ M) failed to significantly block the current ($5.2 \pm 2\%$ at $V_h = +60$ mV, $n = 4$). The failure of capsaicin to activate and capsaizipine to block I_{hTRPV3} agrees with the predicted absence of a capsaicin-binding site²³ in the hTRPV3 sequence.

I_{hTRPV3} activates rapidly in response to voltage steps with no evidence of time-dependent inactivation (Fig. 4a, b). Outside-out, single-channel recordings from hTRPV3-expressing cells revealed that the single channel current–voltage (I – V) relation was outwardly rectifying with a chord conductance of 172 ± 9 pS at +60 mV ($n = 4$; Fig. 4c). At +60 mV, the single-channel mean open time was 5.4 ± 0.2 ms ($n = 4$; Fig. 4d). Consistent with the behaviour of the whole-cell current, single-channel opening probability increased markedly with temperature elevation from 25 °C to

37 °C (Fig. 4e, f).

Activation of I_{hTRPV3} at –60 mV is correlated with the initial rate of heating (Fig. 5a). Rapid heating (4.1 – 6.0 °C s^{-1}) elicited larger hTRPV3 currents (-648.3 ± 237.2 pA at 32 °C) than slower heating protocols (-158.3 ± 33.0 pA at 32 °C). One method used to quantify the temperature sensitivity of ion channels is the 10-degree temperature coefficient (Q_{10}), corresponding to the inverse slope of $\log(I)$ plotted against $1/T$, where T is absolute temperature (Arrhenius plot^{12,24}; Fig. 5b). Q_{10} values for the membrane conductance of relatively ‘temperature-insensitive’ channels is approximately 1.5. In contrast, hTRPV3-transfected cells exposed to slow heating protocols usually showed an initial activation phase ($Q_{10} = 1.9 \pm 0.3$, $n = 16$ cells) followed by a second steep phase ($Q_{10} = 17.3 \pm 3.0$, $n = 19$ cells) corresponding to ~ 11.4 and ~ 54.0 kcal mol^{-1} , respectively. The slope of the second phase is unlike temperature-insensitive conductances, but comparable to data we replotted from I_{TRPV1} , I_{TRPV2} (refs 2, 14) and I_{heat} in neurons^{10,25,26}. The inflection point of these two phases is referred to as the thermal threshold (T_h)^{8,10,26}. In cells that clearly exhibited both a first and second phase, $T_h = 31 \pm 2$ °C ($n = 12$) was close to the threshold for warm-sensitive neurons^{9,27} and warm-activated currents in neurons^{9,10}. Cells exposed to rapid heating typically exhibited a steep initial activation phase ($Q_{10} = 21.6 \pm 4.2$,

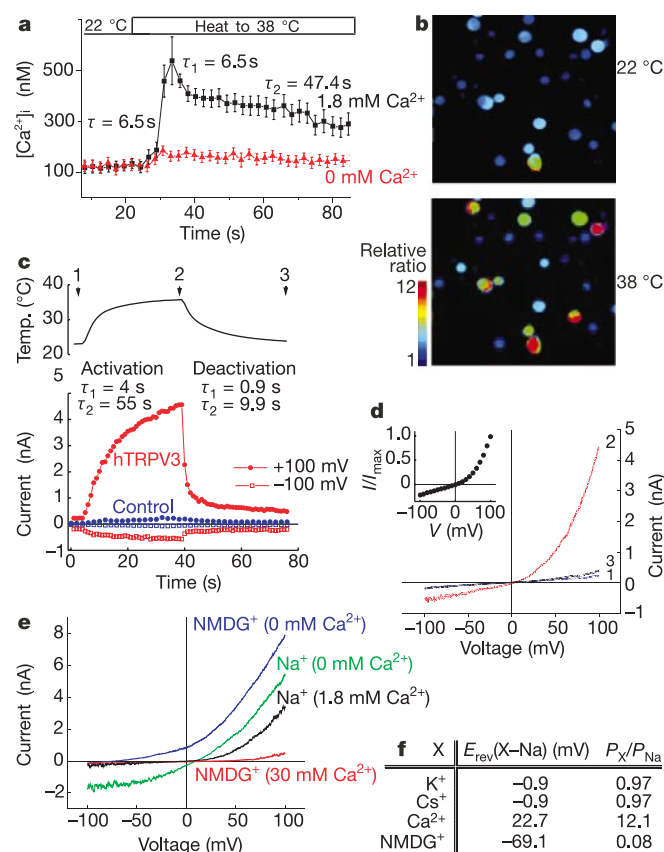


Figure 3 Temperature-activated currents and $[Ca^{2+}]_i$ increases in hTRPV3-expressing CHO-K1 cells. **a**, Fura-2-measured changes of $[Ca^{2+}]_i$ with changing temperature. **b**, Heat-induced Ca^{2+} response in CHO-K1 cells. **c**, Activation of I_{TRPV3} by temperature elevation. Representative whole-cell currents at +100 mV or –100 mV in CHO-K1 cells transfected with hTRPV3 or pTracer vector. **d**, hTRPV3 I – V relation determined from voltage ramps (–100 to +100 mV, 200 ms). Numbers correspond to the time course in **c**. The inset shows normalized I_{hTRPV3} (mean $\pm$ s.e.m., $n = 18$). **e**, I_{hTRPV3} is carried by monovalent cations and Ca^{2+} . **f**, Calculated relative permeabilities of I_{TRPV3} . E_{rev} is the reversal potential; τ is the time constant.

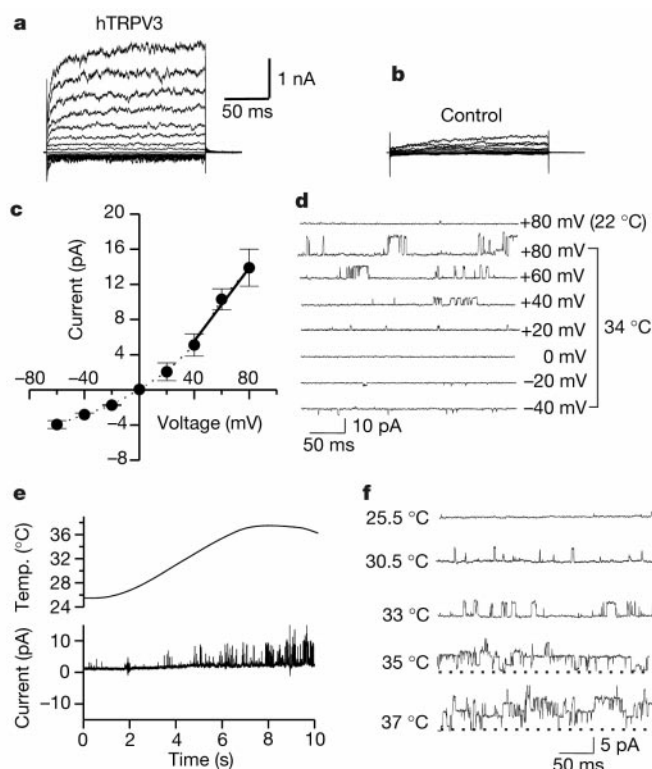


Figure 4 Temperature-activated hTRPV3 currents. **a**, **b**, Kinetics of temperature-activated whole-cell currents resulting from 200-ms voltage steps from –100 to +100 mV in 10-mV increments in CHO-K1 cells transfected with hTRPV3 (**a**) or pTracer vector (**b**). **c**, hTRPV3 single-channel I – V relation at 34 °C. A linear regression fit from +40 to +80 mV (solid line) yielded a slope conductance of 197 ± 11 pS (mean $\pm$ s.e.m., $n = 4$). Chord conductance was 172 ± 9 pS at +60 mV. Channel amplitudes were determined by measuring amplitude histograms at each potential. **d**, Representative single channel traces at different holding potentials. **e**, Single-channel currents (outside-out patches) of hTRPV3-transfected cells in response to a temperature ramp. Current activated most prominently at temperatures above 30 °C. **f**, Representative single-channel currents ($V_h = +60$ mV) in response to a temperature ramp from 24.5 °C to 37 °C. Dashed lines indicate the closed state.

$n = 11$) that was followed by a more shallow phase ($Q_{10} = 3.9 \pm 0.7$, $n = 14$ cells). I_{hTRPV3} did not saturate up to the highest temperature tested (50°C). Similar variable heat responses were also recorded for the capsaicin- and pH-insensitive heat-activated current in trigeminal neurons¹⁰.

Repeated temperature elevation increased both the steady-state current and the current variance (Fig. 5c). hTRPV3 was differentially sensitive to the direction of temperature change, resulting in hysteresis of the thermal activation–deactivation cycle. Whereas increasing temperature resulted in progressive activation of I_{hTRPV3} , even small decreases in temperature profoundly affected current deactivation (Fig. 5d). Loop hysteresis was independent of the absolute temperature reached during the heating protocol (Fig. 5d). Heat-induced activation of I_{hTRPV3} also accelerated the temporal kinetics and decreased the temperature-dependence of subsequent thermal activations (Fig. 5d, e). Q_{10} values for the first, second and third heat exposures in Fig. 5e were similar (18.7, 20.6 and 17.5, respectively). However, the zero current intercept decreased $3\text{--}4^\circ\text{C}$ after the initial heat exposure, indicating a change in the thermal sensitivity of I_{hTRPV3} . The hysteresis effect predicts that the immediate history of temperature activation will affect the current induced in cells expressing hTRPV3. Both the speed of

temperature change and the absolute temperature affect I_{hTRPV3} , consistent with data from primary afferent warm-sensitive fibres²⁷.

We found that hTRPV3 codes for a highly temperature-sensitive ion channel uniquely sensitive around the physiological 37°C temperature point in mammals. The temperature dependence of hTRPV3 fills a gap between the cool-sensing TRPM8 (25°C to 15°C) and the heat-sensing TRPV1 ($>43^\circ\text{C}$) channels. In DRG neurons, not all low-threshold, heat-induced responses are sensitive to capsaicin²⁸. Identification of hTRPV3 as a heat-sensitive but capsaicin-insensitive cation channel may help explain these previous findings. The apparent disparity between robust primate TRPV3 expression reported here and sparse warm-sensitive responses in rodent sensory neurons and fibres⁷ could result from inter-species variability, post-transcriptional or -translational regulation or hetero-oligomerization of TRPV3 with TRPV channels that respond only at higher temperatures. hTRPV3 may report graded responses to warm stimuli in peripheral tissues that are normally below (17°C – 22°C) the channel's thermal threshold, but responds robustly to small perturbations around core body temperature (36°C to 38°C). The wide tissue distribution of hTRPV3 suggests that hTRPV3 may have additional, as yet unidentified functions. □

Methods

Cloning

To identify new TRP channels, we searched the NCBI, Millennium EST, ENSEMBL, and Celera databases with the sequences of known TRPV channels. A full-length human hTRPV3 cDNA containing a 2,373-nucleotide ORF was isolated from a human brain cDNA library using the primers 5'-GCCACGACCACCCAGAACCTCA-3' and 5'-CACCCACTGAGCACTGAGCAGGAG-3'. Both 3' and 5' RACE reactions established that this clone contained the entire ORF.

RT-PCR and *in situ* hybridization

Expression of hTRPV3 in a panel of human nervous system and peripheral tissues was assessed by TaqMan real time quantitative RT-PCR (essentially as previously described²⁹) using the following primers in the 3' end of the coding region, giving an amplicon of 64 bp: sense, 5'-GAGCAGATTCGGATGGGA-3'; antisense, 5'-CCGCAACACAGTCGGA-3'; fluorescent probe, 5'-CATCCTCGGCCACTTTGCACAGCT-3'. β 2-microglobulin RNA was used as the internal reference for gene expression.

To determine cellular expression, 12- μm fresh-frozen sections were prepared from cynomolgus monkey tissues (brain, DRG, spinal cord and SCG). *In situ* hybridization was performed using a ³⁵S-labelled, 247-bp human hTRPV3 cRNA probe covering amino acids 672–754, essentially as described²⁹. A ³³P-labelled cRNA probe (~600 bp) extending from the EcoRI to the EcoRV sites in the 3' end of hTRPV3 was hybridized to a multiple human tissue northern blot (MTN3, Clontech). The same probe labelled with digoxigenin was hybridized to human scalp tissue.

Mammalian cell electrophysiology

hTRPV3 was subcloned into a green fluorescence protein (GFP)-containing vector (pTracer-CMV2, Invitrogen) for expression in CHO-K1 cells. Cells were transfected using LT2 polyamine (PanVera) or Lipofectamine 2000 (Invitrogen) transfection reagent (6–8 h). Transfected cells, cultured at 37°C , were plated onto glass coverslips and recorded 24 to 48 h later. Unless otherwise stated, the pipette solution contained (in mM): 147 Cs, 120 methane-sulphonate, 8 NaCl, 10 EGTA, 2 Mg-ATP, and 20 HEPES, pH 7.4. Bath solution contained (in mM): 138 NaCl, 1.8 CaCl_2 , 5 KCl, 20 HEPES, pH 7.4, and 10 glucose. Data were collected using an Axopatch 2A patch clamp amplifier, Digidata 1320, and pClamp 8.0 software. Whole-cell and excised outside-out patch currents were filtered at 2 kHz and digitized at 10 kHz. The perfusate was heated using a Warner TC-333A or TC-344B heater controller and either Warner SHM-6 or SH-27B solution heater. In some experiments, the bath solution was cooled before heating. The rate of temperature change was between 0.7°C and 6.0°C s^{-1} , depending on the flow rate. For determination of the current–voltage relation, background currents before heat stimuli were subtracted from responses in the presence of heat. In experiments performed with high $[\text{Mg}^{2+}]_i$ to block endogenous currents, the pipette solution contained (in mM): 145 Cs-gluconate, 10 MgCl_2 , 10 $\text{Cs}_4\text{-BAPTA}$, 10 HEPES, pH 7.2. Heat-activated I_{hTRPV3} in high $[\text{Mg}^{2+}]_i$ was not detectably different from those recorded in standard pipette solution.

The permeability of monovalent cations relative to that of Na^+ was estimated from the shift in reversal potential on replacing external Na^+ in nominally Ca^{2+} -free bath solution (145 mM XCl, 20 mM HEPES, 10 mM glucose, pH 7.4), where X was Na^+ , K^+ , Cs^+ , or NMDG⁺. Ca^{2+} permeability was estimated from the shift in reversal potential on replacing NMDG-Cl with solution containing (in mM): 100 NMDG-Cl, 30 CaCl_2 , 20

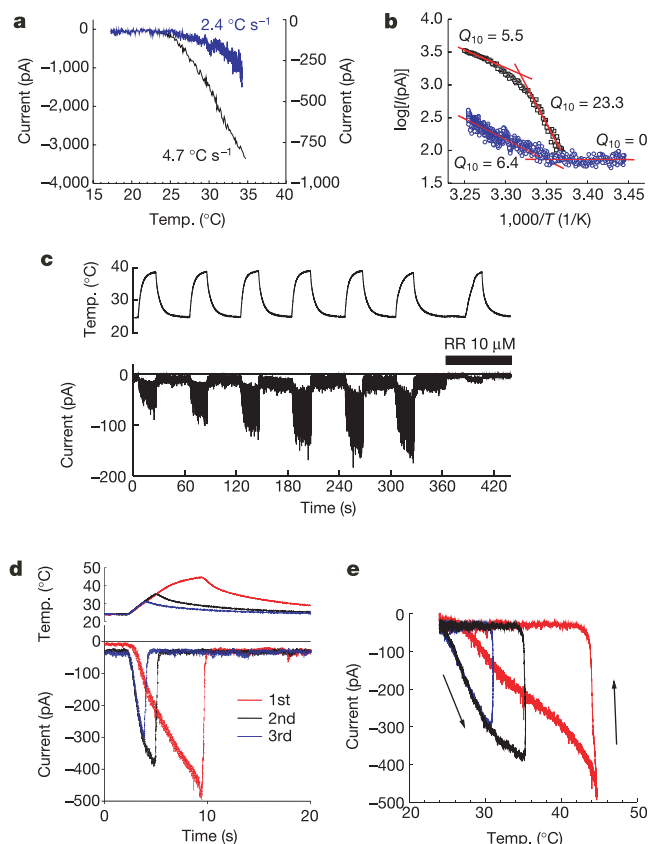


Figure 5 Temperature-dependent gating of I_{hTRPV3} . **a**, I_{hTRPV3} –temperature (T) response from two cells recorded at -60 mV at the initial heating rates indicated. Higher dT/dt (where t is time) characteristically evoked larger currents (black trace, left y axis). **b**, Arrhenius plots of the data in **a** illustrate three phases of the thermal response. Q_{10} values were derived from linear fits of the data and the plots demonstrate the range of variability. **c**, Sensitization of I_{hTRPV3} on repeated heat stimulation is characteristic of I_{hTRPV3} and is blocked by ruthenium red (RR $10\text{ }\mu\text{M}$). **d**, Repeated heating speeds I_{hTRPV3} activation rates without altering deactivation. **e**, Hysteresis of I_{hTRPV3} in response to heating and cooling. Arrows indicate temporal direction of current response.

HEPES, 10 glucose, pH 7.4 (see equation (2)³⁰).

$$P_X/P_{Na} = \{[Na^+]_o/[X]_o\} \{ \exp((F/RT)(V_X - V_{Na})) \} \quad (1)$$

$$P_{Ca}/P_{Na} = \{[Na^+]_o/4[Ca^{2+}]_o\} \{ \exp((F/RT)(V_{Ca} - V_{Na})) \} \{ 1 + \exp(FV_{Ca}/RT) \} \quad (2)$$

where $[X]_o$ is defined as the extracellular concentration of the given ion, P is defined as the permeability of the ion indicated by the subscript, F is Faraday's constant, R is the gas constant, T is absolute temperature and V is the reversal potential for the ion indicated by the subscript.

Calcium imaging

Cells were loaded with 2 μ M Fura-2 AM in cell culture medium at room temperature for 30 min. We recorded Fura-2 ratios (F340/F380) on a MERLIN imaging system (Olympus). The standard curve for Fura-2 ratio versus Ca^{2+} concentration was constructed using the Fura-2 calcium imaging calibration kit (Molecular Probes).

Data analysis

Group data are presented as mean $\pm$ s.e.m. Statistical comparisons were made using analysis of variance and the t -test with Bonferroni correction. A two-tailed value of $P < 0.05$ was taken to be statistically significant.

Received 26 April; accepted 28 May 2002; doi:10.1038/nature00882.

Published online 23 June 2002; corrected 11 July (details online).

- Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
- Caterina, M. J., Rosen, T. A., Tominaga, M., Brake, A. J. & Julius, D. A capsaicin-receptor homologue with a high threshold for noxious heat. *Nature* **398**, 436–441 (1999).
- McKemy, D. D., Neuhauser, W. M. & Julius, D. Identification of a cold receptor reveals a general role for TRP channels in thermosensation. *Nature* **416**, 52–58 (2002).
- Peier, A. M. *et al.* A TRP channel that senses cold stimuli and menthol. *Cell* **108**, 705–715 (2002).
- Spray, D. C. Cutaneous temperature receptors. *Annu. Rev. Physiol.* **48**, 625–638 (1986).
- Reichling, D. B. & Levine, J. D. Heat transduction in rat sensory neurons by calcium-dependent activation of a cation channel. *Proc. Natl Acad. Sci. USA* **94**, 7006–7011 (1997).
- Gotoh, H., Akatsuka, H. & Suto, K. Warm cells revealed by microfluorimetry of Ca^{2+} in cultured dorsal root ganglion neurons. *Brain Res.* **796**, 319–322 (1998).
- Cesare, P., Moriondo, A., Vellani, V. & McNaughton, P. A. Ion channels gated by heat. *Proc. Natl Acad. Sci. USA* **96**, 7658–7663 (1999).
- Hori, A., Minato, K. & Kobayashi, S. Warming-activated channels of warm-sensitive neurons in rat hypothalamic slices. *Neurosci. Lett.* **275**, 93–96 (1999).
- Liu, L. & Simon, S. A. Capsaicin, acid and heat-evoked currents in rat trigeminal ganglion neurons; relationship to functional VR1 receptors. *Physiol. Behav.* **69**, 363–378 (2000).
- Brengelmann, G. L. in *Textbook of Physiology Volume 2: Circulation, Respiration, Body Fluids, Metabolism, and Endocrinology* (eds Patton, H. D., Fuchs, A. F., Hille, B., Scher, A. M. & Steiner, R.) 1584–1596, (W.B. Saunders, Philadelphia, 1989).
- Hille, B. *Ion Channels of Excitable Membranes* (Sinauer Associates, Sunderland, Massachusetts, 2001).
- Möller, S., Croning, M. D. & Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646–653 (2001).
- Tominaga, M. *et al.* The cloned capsaicin receptor integrates multiple pain-producing stimuli. *Neuron* **21**, 531–543 (1998).
- Fundin, B. T., Rice, E. L., Pfaff, K. & Arvidsson, J. The innervation of the mystacial pad in the adult rat studied by anterograde transport of HRP conjugates. *Exp. Brain Res.* **99**, 233–246 (1994).
- Kanzaki, M. *et al.* Translocation of a calcium-permeable cation channel induced by insulin-like growth factor-1. *Nature Cell Biol.* **1**, 165–170 (1999).
- Liedtke, W. *et al.* Vanilloid receptor-related osmotically activated channel (VR-OAC), a candidate vertebrate osmoreceptor. *Cell* **103**, 525–535 (2000).
- Strotmann, R., Harteneck, C., Nunnenmacher, K., Schultz, G. & Plant, T. D. OTRPC4, a nonselective cation channel that confers sensitivity to extracellular osmolarity. *Nature Cell Biol.* **2**, 695–702 (2000).
- Watanabe, H. *et al.* Activation of TRPV4 channels (hVRL-2/mTRP12) by phorbol derivatives. *J. Biol. Chem.* **277**, 13569–13577 (2002).
- Prakriya, M. & Lewis, R. S. Separation and characterization of currents through store-operated CRAC channels and Mg^{2+} -inhibited cation (MIC) channels. *J. Gen. Physiol.* **119**, 487–508 (2002).
- Nadler, M. J. *et al.* LTRPC7 is a Mg-ATP-regulated divalent cation channel required for cell viability. *Nature* **411**, 590–595 (2001).
- Runnels, L. W., Yue, L. & Clapham, D. E. TRP-PLIK, a bifunctional protein with kinase and ion channel activities. *Science* **291**, 1043–1047 (2001).
- Jordt, S. E. & Julius, D. Molecular basis for species-specific sensitivity to 'hot' chili peppers. *Cell* **108**, 421–430 (2002).
- DeCoursey, T. E. & Cherny, V. V. Temperature dependence of voltage-gated H^+ currents in human neutrophils, rat alveolar epithelial cells, and mammalian phagocytes. *J. Gen. Physiol.* **112**, 503–522 (1998).
- Davis, J. B. *et al.* Vanilloid receptor-1 is essential for inflammatory thermal hyperalgesia. *Nature* **405**, 183–187 (2000).
- Vyklicky, L. *et al.* Temperature coefficient of membrane currents induced by noxious heat in sensory neurons in the rat. *J. Physiol.* **517**, 181–192 (1999).
- Andrew, D. & Craig, A. D. Spinothalamic lamina I neurones selectively responsive to cutaneous warming in cats. *J. Physiol.* **537**, 489–495 (2001).
- Nagy, I. & Rang, H. P. Similarities and differences between the responses of rat sensory neurons to noxious heat and capsaicin. *J. Neurosci.* **19**, 10647–10655 (1999).
- Qu, Y. *et al.* Differential modulation of sodium channel gating and persistent sodium currents by the $\beta 1$, $\beta 2$, and $\beta 3$ subunits. *Mol. Cell Neurosci.* **18**, 570–580 (2001).
- Lewis, C. A. Ion-concentration dependence of the reversal potential and the single channel conductance of ion channels at the frog neuromuscular junction. *J. Physiol.* **286**, 417–445 (1979).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank U. Berger, L. Yue, X. Wei, W. Yu, A. Kabakov, L. Runnels, J. Pulido, W. Cao and H. Ferreira for assistance. We are grateful to S. Glucksmann, O. Tayber, L. DeFelice and members of Clapham laboratory for discussion and comments.

Competing interests statement

The authors declare competing financial interests: details accompany the paper on Nature's website (<http://www.nature.com/nature>).

Correspondence and requests for materials should be addressed to D.E.C.

(e-mail: dclapham@enders.tch.harvard.edu). Sequences have been deposited in GenBank under accession number AF514998 (hTRPV3).

TRPV3 is a temperature-sensitive vanilloid receptor-like protein

G. D. Smith[†], M. J. Gunthorpe[†], R. E. Kelsell[‡], P. D. Hayes[‡], P. Reilly^{*}, P. Facer[§], J. E. Wright^{*}, J. C. Jerman^{||}, J.-P. Walhin[‡], L. Ooi^{*}, J. Egerton^{*}, K. J. Charles^{*}, D. Smart^{*}, A. D. Randall^{*}, P. Anand[§] & J. B. Davis^{*}

^{*} Neurology-CEDD, [‡] Genetics Research, and ^{||} Discovery Research, GlaxoSmithKline, Third Avenue, Harlow CM19 5AW, UK
[§] Peripheral Neuropathy Unit, Imperial College, Hammersmith Hospital, Du Cane Rd, London W12 0NN, UK
[†] These authors contributed equally to this work

Vanilloid receptor-1 (VR1, also known as TRPV1) is a thermosensitive, nonselective cation channel that is expressed by capsaicin-sensitive sensory afferents and is activated by noxious heat, acidic pH and the alkaloid irritant capsaicin¹. Although VR1 gene disruption results in a loss of capsaicin responses, it has minimal effects on thermal nociception^{2,3}. This and other experiments—such as those showing the existence of capsaicin-insensitive heat sensors in sensory neurons⁴—suggest the existence of thermosensitive receptors distinct from VR1. Here we identify a member of the vanilloid receptor/TRP gene family, vanilloid receptor-like protein 3 (VRL3, also known as TRPV3), which is heat-sensitive but capsaicin-insensitive. VRL3 is coded for by a 2,370-base-pair open reading frame, transcribed from a gene adjacent to VR1, and is structurally homologous to VR1. VRL3 responds to noxious heat with a threshold of about 39 °C and is co-expressed in dorsal root ganglion neurons with VR1. Furthermore, when heterologously expressed, VRL3 is able to associate with VR1 and may modulate its responses. Hence, not only is VRL3 a thermosensitive ion channel but it may represent an additional vanilloid receptor subunit involved in the formation of heteromeric vanilloid receptor channels.

On the basis of the reasoning that a family of thermoreceptors homologous to VR1 might exist, we searched public nucleotide databases for sequences homologous to those of vanilloid receptor family members. Because we had previously noted that two members of this family, TRPV6/CaT1 (ref. 5) and TRPV5/ECAC1 (ref. 6), were co-localized on chromosome 7 (ref. 7), we focused our searches in genomic regions flanking TRPV1/VR1 17p13 (ref. 8) and TRPV4/VRL-2 12q23–24.1 (ref. 9). Sequence searches of the GenBank database led to the identification of an unfinished human genomic sequence (AC025125) sharing homology with human VR1 (AJ277028). Polymerase chain reaction (PCR) and rapid amplification of complementary DNA ends (RACE) analyses were used to extend the partial gene prediction. Analysis of the sequences

revealed a new vanilloid receptor-like protein, which we refer to as vanilloid receptor-like protein 3 (VRL3), and may be nominated as TRPV3 in the newly defined TRP gene nomenclature¹⁰. A partial, homologous sequence without an ascribed function has been noted previously by ref. 11. Analysis of the longest version of human (h)VRL3 indicated that, similar to hVR1, it is divided into 17 coding exons and one upstream non-coding exon (Fig. 1a). By linking the sequenced tagged site (STS) sequences G54317, G54316, G54321 and G54306 to AC025125—the genomic sequence first identified in our searches—VRL3 was mapped to chromosome 17p13, which is close to VR1 (ref. 12). The most 5' sequence of VRL3 obtained and the most 3' region of VR1 coincided on one fragment of AC027796, placing the two genes only ~7.5 kb apart and in the same transcriptional orientation (Fig. 1a). It remains to be determined whether their transcription is co-regulated; however, the juxtapositioning of VR1 and VRL3 provides further evidence of gene duplication events during the evolution of these channels and raises the likelihood of the genes having similar functional properties⁷.

The complete hVRL3 open reading frame (ORF) of 2,370 bp codes for a polypeptide of 790 amino acids (Fig. 1b) with 43%, 42%, 41% and 30% identity to TRPV1/hVR1 (ref. 8), TRPV4/hVRL-2 (ref. 13), TRPV2/hVRL-1 (ref. 14) and TRPV6/hCaT1 (ref. 5) or TRPV5/hECAC⁶, respectively (see Supplementary Information for alignment). The predicted structure of the protein encompasses a number of features present in VR1 including cytoplasmic termini, ankyrin domains, six transmembrane segments and a pore re-entrant loop, along with several putative phosphorylation sites (Fig. 1b). During amplification of the cDNA, a number of splice variants matching the genomic exon/intron structure were identified; the functional consequences and expression patterns of these splice forms remains to be determined.

VR1 is a polymodal receptor activated by capsaicin, by pH of less than ~6.0 and heat greater than ~42 °C, whereas VRL-1 is activated by heat alone¹⁴. Using electrophysiological and fluorometric techniques applied to HEK293 cells transiently transfected with hVRL3 we examined whether VRL3 possessed a similar activation profile to VR1. In both types of assay VRL3 consistently failed to respond to capsaicin or any other chemical activator of VR1, including extracellular acidification, even at concentrations maximal for VR1 activation (Fig. 2a). In contrast, mild heat challenges evoked currents in VR1 (Fig. 2b) and in approximately 70% of cells transfected with VRL3 (Fig. 2c; $n = 21$). As described previously⁸, untransfected or mock-transfected HEK293 cells only exhibit small 'physicochemically mediated' heat-activated currents (0–40 pA over 25–48 °C, $n = 4$), which are not associated with increased current variance and are linearly related to temperature (Fig. 2b, c). Temperature response curves for VRL3 and VR1 current established that VRL3 has a marked threshold for activation at ~39 °C (Fig. 2c), compared with ~42 °C for VR1 (Fig. 2b), and gives rise to substantial currents at higher temperatures (303 ± 134 pA at 50 °C, $n = 5$) in VRL3 transfected cells. VR1 currents were typically fully activated within ~500 ms, whereas VRL3 currents exhibited slower kinetics and peaked within ~2 s. The VRL3-mediated currents were always associated with an increase in current noise (Fig. 2c), which was analysed (variance analysis) to estimate the current carried by the unitary events. Unitary currents were estimated to be -1.53 ± 0.21 pA and -3.33 ± 0.25 pA ($n = 5$, $P < 0.001$), yielding conductance estimates of 21.9 ± 3.0 pS and 47.6 ± 3.6 pS for VR1 and VRL3 respectively, under identical recording conditions at 48–50 °C. Thus VRL3 has a conductance similar to other TRPV family members¹⁵ but apparently at a level twofold greater than VR1; however, this difference may be modified, such as by the effect of divalent ions.

Current–voltage relationships were established for heat-gated VRL3-mediated currents (Fig. 2d) and found to be similar to that for VR1 (refs 1, 8), exhibiting a reversal potential close to 0 mV ($E_{rev} = -3.8 \pm 2.2$ mV, $n = 3$) and pronounced outward rectifica-

tion ($I_{+70\text{ mV}}/I_{-70\text{ mV}} = 6.8 \pm 1.7$, $n = 3$). Ruthenium red, a non-selective TRPV channel blocker¹⁵, reversibly inhibited ($75 \pm 9\%$ inhibition, $n = 4$) hVRL3 heat-gated currents (Fig. 2e), similar to its effect on VR1 ($79 \pm 5\%$ inhibition, $n = 3$; data not shown). Capsazepine, a tool antagonist of VR1 (ref. 8, 15), did not inhibit VRL3 (Fig. 2f).

We measured VRL3 messenger RNA expression in human tissues by Taqman quantitative PCR (Fig. 3a). VRL3 was abundantly expressed in the central nervous system (CNS) with lower levels in peripheral tissues, suggesting that it is predominantly a gene of the nervous system. Target-specific polyclonal antisera to hVRL3 and hVR1 were used to investigate expression in human dorsal root ganglia (DRG). Both VRL3- and VR1-immunoreactive neurons were scattered throughout control DRG, although VRL3 appeared in fewer cells than VR1. Strong VRL3 immunoreactivity was found

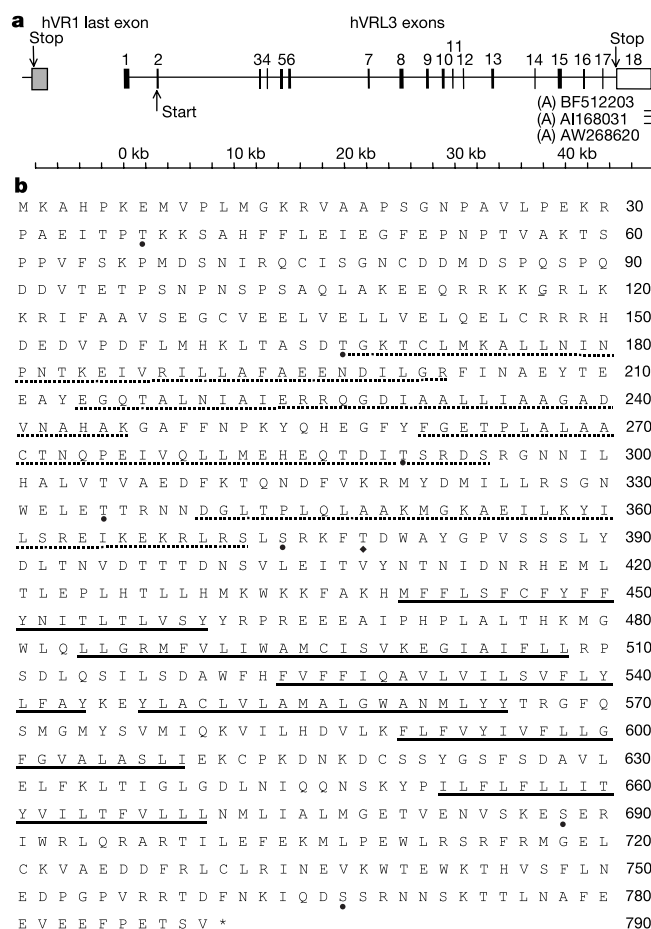


Figure 1 Genomic and protein structure of hVRL3. **a**, Genomic structure and proximity of hVRL3 to hVR1. hVRL3 extends over approximately 48 kb (see scale bar). Exons in black represent experimentally verified sequences. Exon 18 of hVRL3 is extended (indicated by open box) by the expressed sequence tags (ESTs) indicated, which contain a polyadenylation signal also preserved in the genomic sequence (indicated by (A)). The VRL3 translation start and stop sites and last exon of VR1 are indicated three base pairs into exon 2, 93 bp into the last exon and 7.5 kb upstream of hVRL3 (in grey), respectively. **b**, Polypeptide sequence of human VRL3. Amino acid residues are numbered in the margin. Ankyrin repeat domains 1–6 (solid lines) were predicted using SMART²⁷ and TMPRED²⁸; consensus sites for protein kinase A (filled diamond) and protein kinase C (filled circles) are shown. On the basis of the structure of other family members, the pore loop region is presumed to be situated between transmembrane domains 5 and 6.

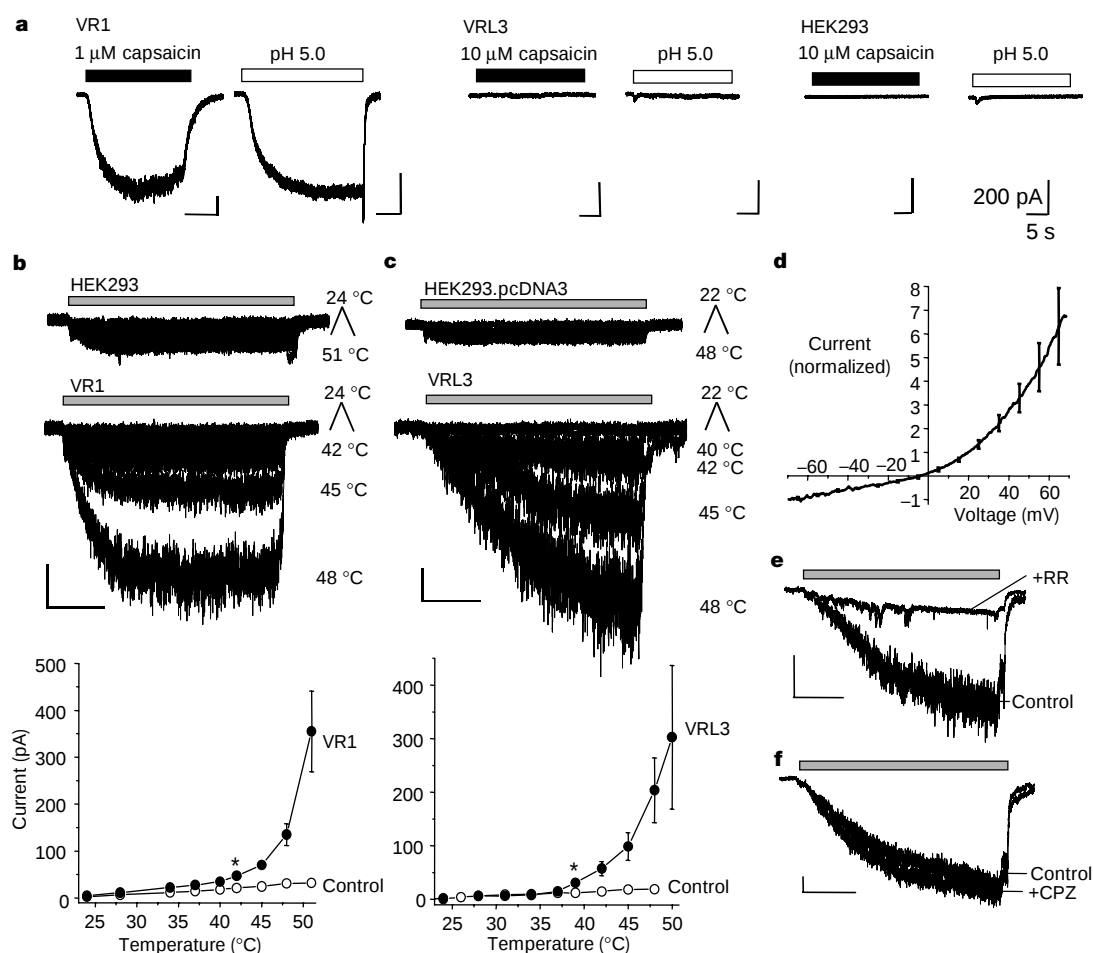


Figure 2 Activation of hVR1 and hVRL3 by noxious heat, capsaicin or protons. **a**, Whole-cell patch clamp recordings of membrane currents in VR1-, VRL3-transfected or untransfected HEK293 cells exposed to capsaicin or pH 5.0. Robust responses were seen with VR1 (1 μ M capsaicin, 803 ± 96 pA, $n = 6$; pH 5.0, 460 ± 83 pA, $n = 6$). **b, c**, Whole-cell patch clamp recordings of membrane currents in control cells (top panels) and hVR1- (**b**) or hVRL3-transfected (**c**) cells (middle panels) recorded in response to the application of extracellular solutions over the temperature range indicated. Temperature-response curves are shown in the bottom panels (filled circles, $n = 4-6$ for all

temperature points shown). Statistical significance (asterisk, $P < 0.05$) at temperatures ≥ 42 °C compared with control for VR1 and ≥ 39 °C compared with control for VRL3 are indicated. **d**, Current-voltage relationship for hVRL3 gated at 50–53 °C. Net current was ascertained by subtraction of control data recorded at 23–25 °C ($n = 3$). **e, f**, Antagonism of VRL3 currents evoked by 50–52 °C heat. Ruthenium red (RR, 10 μ M) blocks ($75 \pm 9\%$ inhibition, $n = 4$) VRL3 currents (**e**); however, capsazepine (CPZ, 10 μ M) does not (**f**). Scale bars: **a**, all 200 pA (vertical), 5 s (horizontal); **b**, 100 pA, 500 ms; **c**, 100 pA, 500 ms; **e**, 250 pA, 500 ms; **f**, 200 pA, 500 ms.

in cells with a diameter $\leq 50 \mu\text{m}$ (Fig. 3b) and much weaker immunoreactivity in a few (about 10% of the total) larger ($>50 \mu\text{m}$) cells. Serial sections showed that VRL3 and VR1 immunoreactivities were localized together in sensory neurons (Fig. 3b, c). VRL3 or VR1 immunoreactivity was completely abolished by preabsorption with their respective antigens (Fig. 3d). In avulsed DRG, the number of cells with diameter $\leq 50 \mu\text{m}$ that were immunoreactive was increased significantly for both VRL3 and VR1 in comparison with age-matched controls (Fig. 3e). Counts of either VRL3 or VR1 immunoreactive cells with a diameter of greater than $50 \mu\text{m}$ showed no change after injury. No correlation was detected between surgical delay or age of subject, or between the number of VRL3 or VR1 immunoreactive cells in control or injured DRG.

One aspect of vanilloid receptor physiology that has not been fully explored to date is the occurrence of heteromeric channel assemblies, as observed for P2X¹⁶ or ASIC¹⁷ ion channels. Their shared chromosomal co-localization, co-expression in neurons, and heat-mediated gating suggests that VR1 and VRL3 may be ideal candidates for intrafamily heteromerization. To test this hypothesis, VRL3 was tagged with an Myc epitope to facilitate immunopreci-

pitiation, and immunoprecipitated from HEK293 cells co-transfected with VR1 and VRL3. VR1 was co-precipitated (Fig. 4a top panel) in a fashion that was dependent on the presence of VRL3 in the immune complex. Similarly, precipitation of VR1 was found to co-precipitate VRL3 (Fig. 4a second panel), which was not due to nonspecific interactions with the supports, Myc tag or precipitating antibody.

Elevating the expression of a channel subunit and changing channel subunit composition may be a control mechanism in afferent activation, possibly operating through alteration of ion channel biophysics and/or pharmacology. To study whether the presence of VRL3 affects the functional properties of VR1, cells were co-transfected with VR1, plus either vector control or VRL3, and the VR1 response then measured by monitoring intracellular calcium. Cells transfected with VR1 alone responded to capsaicin or protons, whereas cells transfected with VRL3 alone failed to respond to either ligand (Fig. 4b, c). Co-expression of VRL3 with VR1 produced a substantial increase in the response to capsaicin (Fig. 4b) and a similar increase in the response to protons (Fig. 4c). Capsazepine was able to block the response of VR1 and VRL3 co-transfected cells to capsaicin or protons with no increase in

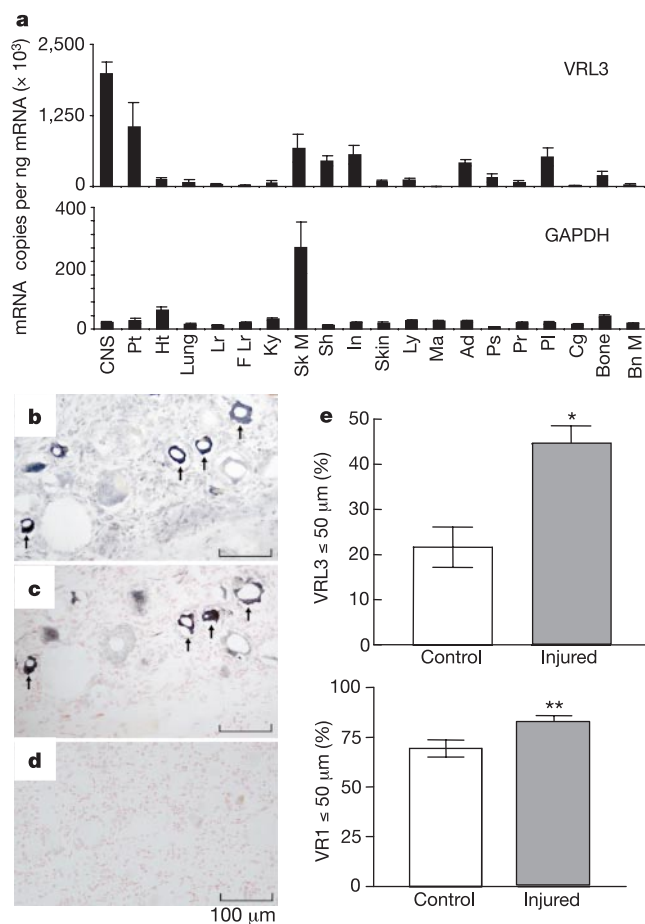


Figure 3 Localization of VRL3. **a**, mRNA expression of hVRL3 in human tissues. Copies of hVRL3 (top) or GAPDH (bottom) per 50 ng total RNA ($n = 4$, mean $\pm$ s.e.m.). Pt, pituitary; Ht, heart; Lr, liver; F Lr, fetal liver; Ky, kidney; Sk M, skeletal muscle; Sh, stomach; In, intestine; Ly, lymphocytes; Ma, macrophages; As, adipose; Ps, pancreas; Pr, prostate; Pl, placenta; Cg, cartilage; Bn M, bone marrow. **b, c**, Serial, human DRG sections immunostained with antibodies to hVRL3 (**b**) or hVR1 (**c**) showing the same cells immunoreactive for both antigens (arrows). **d**, DRG sections immunostained with antibodies pre-incubated with homologous antigen to hVRL3. Scale bars, 100 μ m. **e**, Quantification of VRL3- and VR1-immunoreactivity in neurons from control or injured DRG. Bar charts show a significant increase in the number of immunoreactive cells after injury. VRL3: asterisk, $P = 0.0025$; VR1: double asterisk, $P = 0.015$ (Student's t -test).

residual activity compared to VR1 transfected cells, suggesting that capsazepine is equally able to block any VR1–VRL3 heteromeric channels formed. This data suggests that VRL3 may contribute to VR1 responses by a mechanism that might, given the evidence for interaction described above, entail formation of heteromeric channels. At this stage, however, we cannot rule out that the functional effect might occur as a result of alterations in VR1 transport or competition for desensitizing mechanisms.

Although additional TRPV family members have been identified, none of the functional properties of the previously identified paralogues provide a complete explanation for heat responses in capsaicin-insensitive neurons⁴ or the thermal nociception remaining in VR1 knockout mice^{2,3}. In contrast to TRPV2/VRL-1 (heat threshold $\sim 52^\circ\text{C}$), VRL3 responds in a similar temperature range to VR1 and is a candidate for the ‘unexplained’ heat sensor of VR1 knockout mice. It is unclear then why VRL3 activity was not detected in the neonatal sensory neurons of VR1 knockout mice^{2,3}. This may be a consequence of the age of ganglia dissected or the chosen culture conditions, as previously suggested³. Another possibility is that the VR1 knockout strategies have deleted

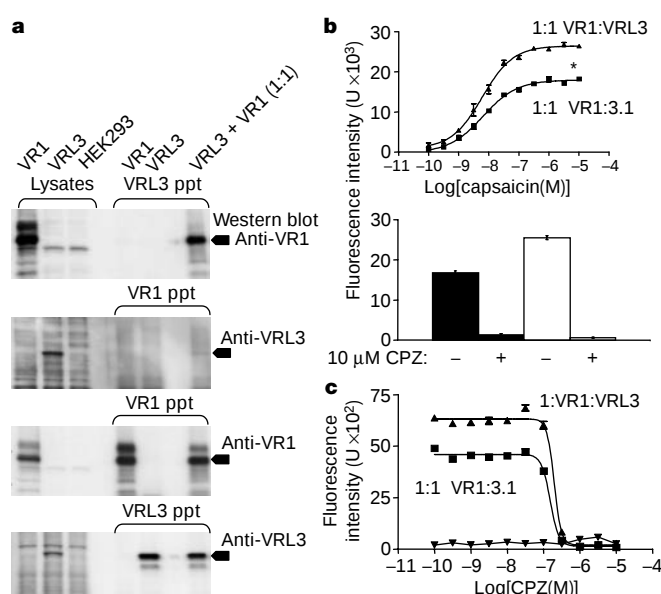


Figure 4 Association between VR1 and VRL3. **a**, Co-precipitation of VR1 and VRL3. Lysates from HEK293 cells co-transfected with VR1 and VRL3-Myc or vector were analysed by western blot before or after affinity purification using C22 or anti-Myc antibodies to precipitate VR1 or VRL3-Myc, respectively. VR1 co-precipitated with VRL3-Myc (top panel) and hVRL3-Myc co-precipitated with VR1 (second panel, shown by arrows). **b**, Enhancement of VR1 response to capsaicin by co-expression with VRL3 (upper panel). HEK293 cells co-transfected with VR1 plus vector control (squares) or VR1 plus hVRL3 (triangles) were challenged with the capsaicin concentrations shown (asterisk, $P < 0.0001$, $n = 4$). Capsazepine (CPZ, 10 μ M) antagonized the response to 100 nM capsaicin of cells co-transfected with VR1 plus vector control (black bars) or VR1 plus VRL3 (white bars, bottom panel). **c**, The enhanced response of VR1 (squares) to protons (pH 5.0) when co-transfected with VRL3 (triangles) and the inhibition of the response by CPZ 10 μ M are shown. Changes in intracellular calcium were expressed as mean change of peak minus basal fluorescence intensity $\pm$ s.e.m. ($n = 4$). ppt, precipitation.

transcriptional control elements shared between VR1 and VRL3. These considerations lead to the possibility that VRL3 disruption contributed to the phenotype ascribed to disruption of VR1 and also to the possible existence of yet further thermal receptors—questions which may be answered by VRL3 and VR1 knockout studies.

The observations of co-expression of VRL3 and VR1 in individual sensory neurons and the biochemical evidence for association between the two proteins, gives rise to a hypothesis that VRL3, or possibly other TRP proteins, may contribute subunits to vanilloid receptor-like ion channels. Such heteromerization may allow sensory neurons to express a range of receptors with a spectrum of response characteristics wider than those of the individual homomeric receptors. Thus, the identification of VRL3 and its different threshold and kinetic characteristics compared to VR1 supports the view that members of the TRPV family may constitute a family of thermal sensors that code for a dynamic range of thermal sensitivity in sensory neurons¹⁴. A further development, however, is that differential regulation of TRPV expression, or the formation of heteromeric interactions, may underlie the thermal hypersensitivity of sensory neurons at sites of tissue injury and inflammation. □

Methods

Cloning of cDNA for hVRL3

Homology searching and gene predictions were performed using TBLASTN2 (ref. 18), GENEWISE¹⁹ and GENSCAN²⁰ programs. All other general sequence manipulations were performed with GCG version 9 and 10.2 (Genetics Computer Group Inc). We conducted

cDNA cloning using previously described methods⁸. Primers derived from the *hVRL3* gene prediction sequences were used to amplify cDNAs corresponding to the central part and the 3' end by PCR with reverse transcription (RT-PCR) from brain, heart and testis poly(A)⁺ RNA (Clontech). RACE was performed with small intestine, colon and bladder poly(A)⁺ RNAs using the SMART RACE kit (Clontech). It was necessary to include 10% DMSO in the RACE reactions because of the high G + C content of the 5' end. The complete coding region of the *hVRL3* cDNA was amplified from small intestine and colon cDNAs by nested PCR using, for the primary reaction, forward 5'-TTGCCACGACCA CCCAGAAC-3' and reverse 5'-CAGACGGCTGCTGGCTGTAG-3' primers, and for the secondary reaction, forward 5'-ACCATGAAGCCACCCCAAG-3' and reverse 5'-TCTCTGCACAGATCGGTGAC-3' primers. Multiple full-length cDNA clones, from independent PCR reactions, were double-strand-sequenced to rule out PCR-introduced mutations.

Electrophysiology

HEK293 cells were cultured and transiently transfected with pcDNA3.1-hVR1 or pcDNA3.1-hVRL3 and pIRES-GFP, prepared on coverslips, and whole-cell patch-clamp recordings were performed as previously described⁸. Experiments were conducted at room temperature (20–24 °C) unless otherwise stated. The extracellular solution consisted of (in mM): NaCl 130, KCl 5, CaCl₂ 2, MgCl₂ 1, glucose 30, HEPES-NaOH 25, pH 7.3. Patch pipettes (resistance 2–5 MΩ) were filled with the following solution (in mM): CsCl 140, MgCl₂ 4, EGTA 10, HEPES-CsOH 10, pH 7.3. Application of drugs and temperature jumps were applied using an automated device for fast switching of solutions coupled with an in-line heating device (Warner Instruments in-line heater SH-27A) as described⁸. Antagonists were pre-applied for at least 30 s. Data were acquired at 4 kHz and filtered at 2 kHz, and analysis was performed using the pClamp7 software suite and Origin (Microcal). Variance (noise) analysis was conducted as described²¹. All data are presented as the mean ± s.e.m.

Calcium imaging

Changes in intracellular calcium were monitored using a fluorometric imaging plate reader (FLIPR) platform as previously described²². Briefly, HEK293 cells transiently transfected with pcDNA3.1-hVRL3, pcDNA3.1-hVR1, vector alone or combinations of vectors, were plated into black-walled microtitre plates, loaded with the calcium indicator dye Fluo3 AM, and assayed as described²². Data are expressed as mean ± s.e.m. unless otherwise stated. Curve-fitting and parameter estimation were carried out using Graph Pad Prism 3.00 (GraphPad Software Inc). Statistical comparisons were made where appropriate using Student's *t*-test.

Localization studies

Tissue expression of human VRL3 was studied using Taqman quantitative RT-PCR²³ with 5'-CAAGGTCATCCATGACAACCTTTG-3' and 5'-GGGCCATCCACAGTCTCTTG-3' primers in combination with a 5'-ACCACAGTCCATGCCATCACTGCCA-3' probe for GAPDH, or 5'-CCTCCTCAACATGCTCATGTCT-3' and 5'-ATGCGTTCGCTCT CCTTGG-3' primers with a 5'-CGTTCTCCACAGTCTCGCCCATCA-3' probe for VRL3, as previously described²⁴. Data is presented as the mean number of copies per 50 ng of total RNA from four patients ± s.e.m. Samples were provided by the Netherlands Brain Bank. For immunohistochemical studies, cervical DRG whose roots had been avulsed from the spinal cord after traumatic injury to the brachial plexus were collected from ten adult patients during surgical repair—the delay between injury and collection of DRG at operation ranged from 1 to 74 days. Matched control cervical DRG were obtained through the Netherlands Brain Bank from six subjects with an autopsy delay of less than 12 h. All non-autopsy tissues were removed as a necessary part of surgical repair and not for research purposes; the patients gave informed consent for this study, which had local ethics committee approval. Tissues were stored, sectioned, post-fixed, counterstained and immunoreactivity detected as previously described²⁵. Rabbit polyclonal antibodies C22 and N63, raised to a carboxy terminus peptide from human VR1 ((C)KPEDAIEVFKSPAASGEK) and an amino-terminal peptide from VRL3 (LAKEEQRRKKG(C)), respectively, were used as primary antibodies. For co-localization, serial, inverted 'mirror image', frozen sections were collected and treated in the same fashion. Negative controls included omission of primary antibodies, replacement with pre-immune serum or pre-adsorption of antibodies with antigen. Nucleated, immunoreactive and non-immunoreactive DRG neurons were counted (140 ± 24 per control or 94 ± 14 per injured patient stained for VRL3; 170 ± 8 per control or 154 ± 14 per injured patient stained for VR1) and an approximation of cell diameter obtained using a calibrated microscope eyepiece graticule at × 20 objective magnification. Counts were made by two independent observers.

Immunoprecipitation

An Myc tag was placed at the C terminus of VRL3 using standard cDNA cloning procedures. HEK293 cells were transfected with VR1 or VRL3-Myc as described⁸. Transfected cells were lysed in 2 ml RIPA buffer (20 mM sodium phosphate pH 7.4, 500 mM NaCl, 0.1% SDS, 1% NP-40, 0.5% sodium deoxycholate) containing protease inhibitors (Complete, protease inhibitor cocktail tablets; Roche). Protein complexes were precipitated from pre-cleared lysates of equivalent protein content overnight at 4 °C using anti-Myc clone JAC6 (Serotec) (4 µg) or anti-VR1 antiserum (C22) (4 µg) and protein-A-Sepharose beads, washed with ice cold RIPA buffer and analysed by SDS-polyacrylamide gel electrophoresis and western blot essentially as described²⁶. VR1 or VRL3-Myc were detected using antiserum C22 (1:5,000) or anti-Myc-horseradish peroxidase (HRP) (Invitrogen) and HRP-conjugated donkey anti-rabbit (Sigma) where necessary, followed by detection by chemiluminescence. Myc-tagged proteins unrelated to VR1, such as neuropilin and human γ-2 calcium channel subunit were used as

negative controls to show that association did not occur through the Myc tag. Furthermore, precipitation did not occur when lysates from singly transfected cells were mixed before precipitation, indicating that the VR1-VRL3 association must occur in the cell.

Received 21 February; accepted 29 May 2002; doi:10.1038/nature00894.

Published online 23 June 2002.

1. Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
2. Caterina, M. J. *et al.* Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science* **288**, 306–313 (2000).
3. Davis, J. B. *et al.* Vanilloid receptor-1 is essential for inflammatory thermal hyperalgesia. *Nature* **405**, 183–187 (2000).
4. Nagy, I. & Rang, H. P. Similarities and differences between the responses of rat sensory neurons to noxious heat and capsaicin. *J. Neurosci.* **19**, 10647–10655 (1999).
5. Peng, J. B. *et al.* Human calcium transport protein CaT1. *Biochem. Biophys. Res. Commun.* **278**, 326–332 (2000).
6. Müller, D. *et al.* Molecular cloning, tissue distribution, and chromosomal mapping of the human epithelial Ca²⁺ channel (ECAC1). *Genomics* **67**, 48–53 (2001).
7. Müller, D. *et al.* Gene structure and chromosomal mapping of human epithelial calcium channel. *Biochem. Biophys. Res. Commun.* **275**, 47–52 (2000).
8. Hayes, P. D. *et al.* Cloning and functional expression of a human orthologue of rat vanilloid receptor-1. *Pain* **88**, 205–215 (2000).
9. Delany, N. S. *et al.* Identification and characterization of a novel human vanilloid receptor-like protein, VRL-2. *Physiol. Genom.* **4**, 165–174 (2001).
10. Montell, C. *et al.* A unified nomenclature for the superfamily of TRP cation channels. *Mol. Cell* **9**, 229–231 (2002).
11. Peng, J. B. *et al.* Structural conservation of the genes encoding CaT1, CaT2 and related cation channels. *Genomics* **76**, 99–109 (2001).
12. Touchman, J. W. *et al.* The genomic region encompassing the nephropathic cystinosis gene (CTNS): complete sequencing of a 200 kb segment and discovery of a novel gene within the common cystinosis-causing deletion. *Genome Res.* **10**, 165–173 (2000).
13. Liedtke, W. *et al.* Vanilloid receptor-related osmotically activated channel (VR-OAC), a candidate vertebrate osmoreceptor. *Cell* **103**, 525–535 (2000).
14. Caterina, M. J. *et al.* A capsaicin-receptor homologue with a high threshold for noxious heat. *Nature* **398**, 436–441 (1999).
15. Gunthorpe, M. J., Benham, C. D., Randall, A. & Davis, J. B. The diversity in the vanilloid (TRPV) receptor family of ion channels. *Trends Pharmacol. Sci.* **23**, 183–191 (2002).
16. Liu, M. *et al.* Coexpression of P2X(3) and P2X(2) receptor subunits in varying amounts generates heterologous populations of P2X receptors that evoke a spectrum of agonist responses comparable to that seen in sensory neurons. *J. Pharmacol. Exp. Ther.* **296**, 1043–1050 (2001).
17. Babinski, K., Catarsi, S., Biagini, G. & Seguela, P. Mammalian ASIC2a and ASIC3 subunits co-assemble into heteromeric proton-gated channels sensitive to Gd³⁺. *J. Biol. Chem.* **275**, 28519–28525 (2000).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
19. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
20. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
21. Dempster, J. *Computer Analysis of Electrophysiological Signals* (Academic, London, 1993).
22. Smart, D. *et al.* The endogenous lipid anandamide is a full agonist at the human vanilloid receptor (hVR1). *Br. J. Pharmacol.* **129**, 227–230 (2000).
23. Gibson, U. E., Heid, C. A. & Williams, P. M. A novel method for real time quantitative RT-PCR. *Genome Res.* **6**, 995–1001 (1996).
24. Moore, D. J. *et al.* Expression pattern of human P2Y receptor subtypes: a quantitative reverse transcription-polymerase chain reaction study. *Biochim. Biophys. Acta* **1521**, 107–119 (2001).
25. Coward, K. *et al.* Immunolocalization of SNS/PN3 and Na⁺/SNS2 sodium channels in human pain states. *Pain* **85**, 41–50 (2000).
26. Gray, C. W. *et al.* Characterisation of human HtrA2, a novel serine protease involved in the mammalian cellular stress response. *Eur. J. Biochem.* **267**, 5699–5710 (2000).
27. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: identification of signalling domains. *Proc. Natl Acad. Sci. USA* **95**, 5857–5864 (1998).
28. Hofmann, K. & Stoffel, W. TMbase—a database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler* **374**, 166 (1993).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements

The authors acknowledge R. Birch and R. Ravid for supply of human adult tissues; I. Gloger, C. Benham, M. Duckworth and C. Bountra for advice and encouragement; and C. Farrant for help with artwork.

Competing interests statement

The authors declare competing financial interests: details accompany the paper on *Nature's* website (<http://www.nature.com/nature>).

Correspondence and requests for materials should be addressed to J.B.D. (e-mail: John_B_Davis@gsk.com). The GenBank accession number for human VRL3 is AJ487035.

Release of chromatin protein HMGB1 by necrotic cells triggers inflammation

Paola Scaffidi*, Tom Misteli† & Marco E. Bianchi‡

* DIBIT, Istituto Scientifico San Raffaele, 20132 Milano, Italy

† National Cancer Institute, NIH, Bethesda, Maryland 20892, USA

‡ Università Vita Salute San Raffaele, 20132 Milano, Italy

High mobility group 1 (HMGB1) protein is both a nuclear factor and a secreted protein. In the cell nucleus it acts as an architectural chromatin-binding factor that bends DNA and promotes protein assembly on specific DNA targets^{1,2}. Outside the cell, it binds with high affinity to RAGE (the receptor for advanced glycation end products)³ and is a potent mediator of inflammation^{4–6}. HMGB1 is secreted by activated monocytes and macrophages⁴, and is passively released by necrotic or damaged cells^{7–9}. Here we report that *Hmgb1*^{−/−} necrotic cells have a greatly reduced ability to promote inflammation, which proves that the release of HMGB1 can signal the demise of a cell to its neighbours. Apoptotic cells do not release HMGB1 even after undergoing secondary necrosis and partial autolysis, and thus fail to promote inflammation even if not cleared promptly by phagocytic cells. In apoptotic cells, HMGB1 is bound firmly to chromatin because of generalized underacetylation of histone and is released in the extracellular medium (promoting inflammation) if chromatin deacetylation is prevented. Thus, cells undergoing apoptosis are programmed to withhold the signal that is broadcast by cells that have been damaged or killed by trauma.

HMGB1 is bound loosely to the chromatin of both interphase and mitotic cells, and is leaked rapidly into the medium when membrane integrity is lost in permeabilized or necrotic cells^{7–9}. These results suggest that in living cells HMGB1 associates and dissociates rapidly from chromatin. To prove this, we tagged HMGB1 at its carboxy terminus with green fluorescent protein (GFP), which formed a chimaeric protein that was equivalent to wild-type HMGB1 in enhancing the expression of a HOXD9-responsive reporter gene in transfection assays (ref. 10 and data not shown). HeLa cells expressing the fusion protein were easily detectable by the uniform green fluorescence of their nuclei. Cells undergoing mitosis showed a diffuse cytoplasmic fluorescence, but also a distinct association of HMGB1–GFP with condensed chromosomes that lasted throughout M phase (Fig. 1a and data not shown).

We permeabilized HeLa cells transfected with HMGB1–GFP with Nonidet-P40 (NP-40). Most cells lost their fluorescence after a few seconds, confirming the loose association of HMGB1 with chromatin; however, a few cells retained a bright fluorescence. From the characteristically fragmented appearance of their nuclei, these cells seemed to be apoptotic. We forced HeLa cells to undergo apoptosis by treatment with tumour-necrosis factor- α (TNF- α) and cycloheximide, permeabilized them, and then immunostained them for endogenous, unmodified HMGB1. Although control non-apoptotic cells leaked HMGB1 into the medium (Fig. 1b, d), the protein was retained within the nucleus of apoptotic cells (Fig. 1d). HMGB1 was mostly retained with nuclear remnants even after prolonged incubation and partial autolysis of apoptotic cells, when soluble cytoplasmic proteins such as lactate dehydrogenase leaked into the extracellular medium (Fig. 1e). HMGB1 and HMGB1–GFP also bound tightly to chromatin in HeLa and 3T3 cells that were either induced into apoptosis by treatment with etoposide or H₂O₂, or apoptosing spontaneously in unperturbed cultures (data not

shown). By contrast, HMGB1 dissociated from the chromatin of necrotic cells and leaked into the extracellular medium (Fig. 1c).

We used fluorescence loss in photobleaching (FLIP)¹¹ to quantify the dynamic properties of HMGB1–GFP in single cells. In FLIP, repeated bleaching of the same area leads to fluorescence loss from the rest of the nucleus, with kinetics that are dependent on the overall mobility of the fluorescent protein. If a fraction of the protein pool is at any given time bound to chromatin, the loss of its fluorescence will be slowed. Bleaching of total nuclear HMGB1–GFP was obtained rapidly in living cells (Fig. 2a). By contrast, in HeLa cells that expressed GFP fusions of chromatin proteins HMGN1 and HMGN2, or of transcription factor NF1, fluorescence loss was significantly slower, and in cells expressing GFP–histone H1c fusions fluorescence loss was very limited (ref. 12 and Fig. 2c).

We also assessed the diffusion rate of HMGB1–GFP associated with condensed chromosomes in living HeLa cells during mitosis. The repeated bleaching of cytoplasmic HMGB1–GFP led to a rapid and parallel loss of fluorescence from condensed chromosomes and from the cytoplasm (Fig. 2b, d), which proves unequivocally that

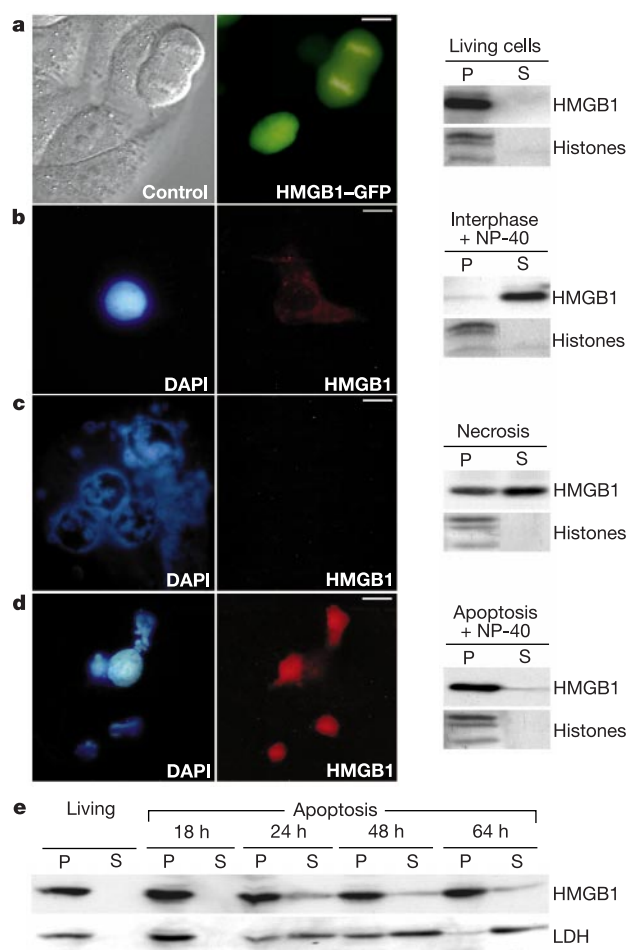


Figure 1 Chromatin association of HMGB1 in living and dead HeLa cells. Both the medium bathing the cells (S) and the cells (P) were analysed by SDS–polyacrylamide gel electrophoresis. Histones were visualized by Coomassie blue staining, HMGB1 by immunoblotting or immunostaining with antibody to HMGB1, DNA by DAPI. Scale bars, 7.5 μ m. **a**, Living cells expressing HMGB1–GFP, imaged by differential interference contrast and in green fluorescence. **b**, Interphase cells after permeabilization. **c**, Necrotic cells with no permeabilization. The amount of HMGB1 in the medium was proportional to the number of necrotic cells (about 50%). **d**, Apoptotic cells with permeabilization. **e**, Kinetics of HMGB1 and lactate dehydrogenase (LDH) release from cells undergoing apoptosis and secondary necrosis.

HMGB1 turns over rapidly between its chromatin-bound and soluble states. At the other extreme, HMGB1-GFP seemed to be almost immobile in apoptotic cells (Fig. 2e, g). The blockade of HMGB1 was specific, because the mobility of GFP-HMGN1, GFP-HMGN2, GFP-NF1 or GFP alone was not reduced in apoptotic cells, as compared with living ones (Fig. 2h and data not shown). Thus, chromatin condensation during apoptosis does not impair protein mobility in general.

We used *Hmgb1*^{-/-} cells to test whether the binding of HMGB1 to apoptotic chromatin was caused by alterations in either HMGB1 or nuclei undergoing apoptosis. Embryonic fibroblasts obtained from *Hmgb1*^{-/-} and *Hmgb1*^{+/+} mice¹³ were equally susceptible to apoptosis (data not shown), which indicated that the freezing of HMGB1 onto chromatin was a consequence of apoptosis but not a

requisite. We treated *Hmgb1*^{-/-} fibroblasts with TNF- α and cycloheximide, and recovered apoptotic cells from the flask by gentle flushing. This cell population, and a control population of non-apoptotic *Hmgb1*^{-/-} fibroblasts, was permeabilized with detergent and exposed to bacterially produced, Cy5-labelled HMGB1. HMGB1 bound to apoptotic nuclei, but not to non-apoptotic ones (Fig. 3a). We verified this result biochemically (Fig. 3b). Permeabilized apoptotic and non-apoptotic *Hmgb1*^{-/-} fibroblasts were incubated with bacterially produced HMGB1 and fractionated through a discontinuous sucrose gradient⁹. Again, HMGB1 associated with the nuclei from apoptotic cells, but not with those from non-apoptotic cells. Together, these findings indicate that on apoptosis chromatin undergoes some chemical or structural transition that makes it susceptible to HMGB1 binding. The nature of

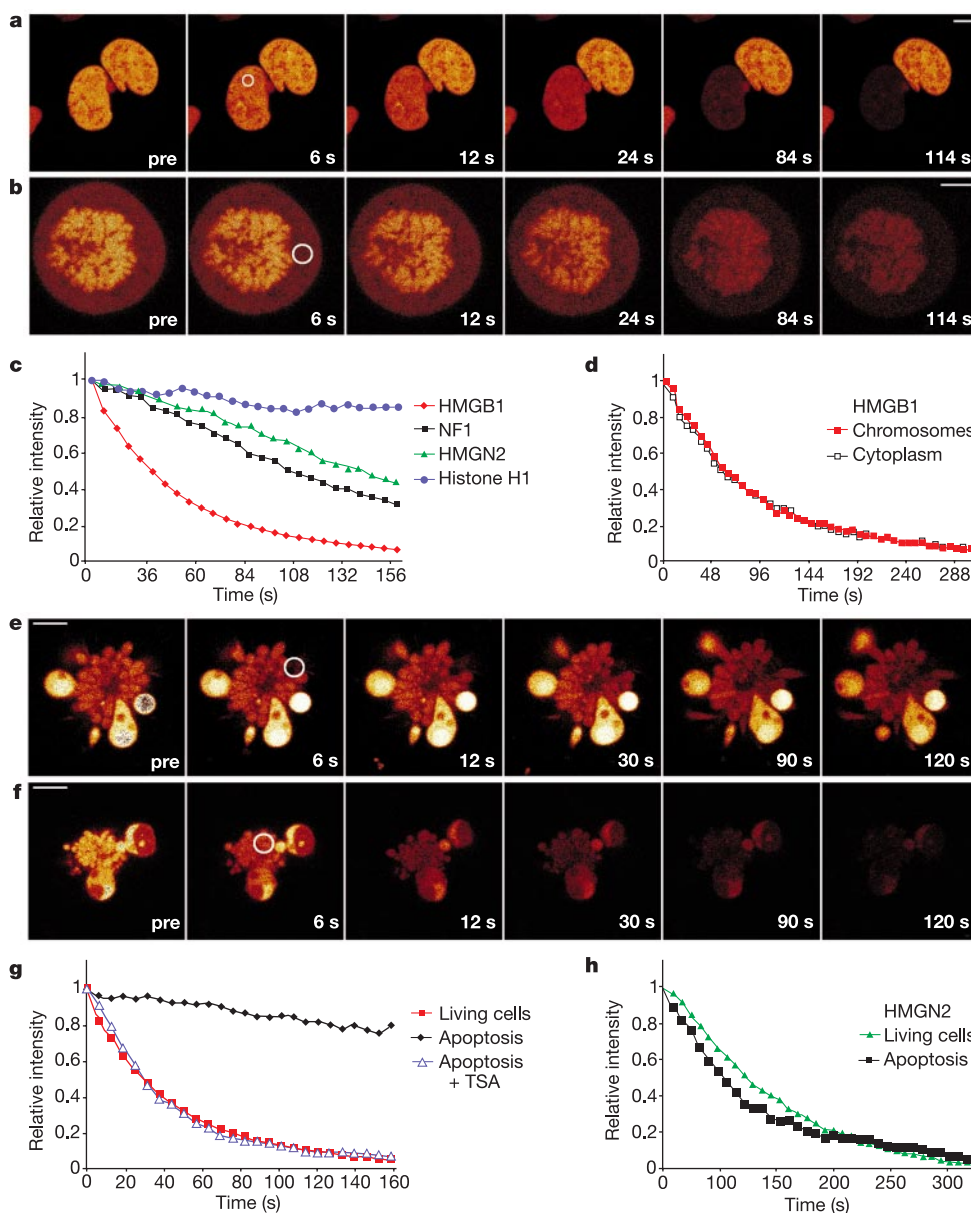


Figure 2 HMGB1 dynamics in living and apoptotic cells. **a, c**, FLIP imaging (**a**) and quantification (**c**) of HMGB1-GFP in an interphase cell. The circled area was bleached repeatedly, and cells were imaged between bleach pulses. A neighbouring cell nucleus was not affected. **b, d**, FLIP imaging (**b**) and quantification (**d**) on a mitotic cell. Bleaching was executed in the cytoplasm (circle), and quantification was done on a different spot in

the cytoplasm or on a spot on the condensed chromosomes. **e-g**, FLIP imaging in an apoptotic cell (**e**) and a cell undergoing apoptosis in the presence of 200 ng ml⁻¹ TSA (**f**), and their quantification (**g**). **h**, FLIP quantification of GFP-HMGN2 in living and apoptotic cells. Scale bars, 2.3 μm (**a, b**) and 3.7 μm (**e, f**).

HMGB1 itself, whether endogenous or produced in bacteria, tagged with fluorophores or fused to GFP, is irrelevant.

We next investigated the nature of the chromatin modification that allows the stable binding of HMGB1. Because HMGB1 binds tightly to *in vitro* reconstructed mononucleosomes^{9,14,15}, we tested whether the fragmentation of chromatin to oligo- and mononucleosomes that occurs in the later stages of apoptosis can provide stable binding sites for HMGB1. HeLa cells were stably transfected with a construct expressing ICAD, the inhibitor of the CAD nuclease that fragments DNA during apoptosis. Although HeLa cells over-expressing ICAD underwent apoptosis, their DNA showed little if any fragmentation (ref. 16 and Fig. 3c). HMGB1 bound equally stably to ICAD-expressing, non-fragmented chromatin and to fragmented chromatin (Fig. 3c and data not shown). DNA fragmentation therefore cannot account for stable HMGB1 binding in apoptosis.

We also tested for alterations in the acetylation status of chromatin. Trichostatin A (TSA), a general deacetylase inhibitor, was added to the medium of HeLa cells just before the induction of apoptosis, which suppressed the binding of HMGB1 on chromatin (Fig. 2f, g). This result suggests that hypoacetylation of one or more chromatin components occurs during apoptosis and favours HMGB1 binding. No difference was seen in the isoelectric point (pI) or the molecular weight pattern of HMGB1 present in apoptotic and non-apoptotic cells (Fig. 3d), which indicates that apoptosis does not change the acetylation status of HMGB1 itself. By contrast, histone H4 from apoptotic chromatin was hypoacetylated in comparison to non-apoptotic chromatin, and H4 hypoacetylation in apoptosis was suppressed by TSA (Fig. 3e and data not shown).

Thus, HMGB1 binding to chromatin depends on the viability of the cell and clearly distinguishes necrotic from apoptotic cells. We therefore reasoned that the differential release of HMGB1 might be exploited as a cue to nearby cells to activate the appropriate responses to unprogrammed and programmed cell death. Unprogrammed death is usually the result of trauma, poisoning or infection, each of which requires prompt reaction, damage containment and/or damage repair. Inflammation is the primary tissue damage response in mammals, and HMGB1 has been reported to be a mediator of inflammation⁴⁻⁶. To test directly whether the release of HMGB1 by necrotic cells might be an immediate trigger for an inflammatory response, we challenged wild-type bone marrow cells with *Hmgbl*^{-/-} or wild-type dead fibroblasts. As expected¹⁷, wild-type necrotic cells triggered the production of the proinflammatory cytokine TNF- α , whereas wild-type apoptotic cells were much less effective (Fig. 4a). Significantly, *Hmgbl*^{-/-} necrotic cells were also ineffective in activating monocytes. Purified HMGB1 also elicits TNF- α production in this assay (ref. 6 and data not shown). Thus, HMGB1 is one of the main diffusible signals of necrosis.

The previous experiment does not prove that apoptotic cells escape the inflammatory surveillance owing to retention of HMGB1, because apoptotic cells start to leak cellular components only after several hours and *in vivo* they are routinely cleared by phagocytic cells long before this process (termed 'secondary necrosis') can take place. But we tested whether cells undergoing post-apoptotic, secondary necrosis can promote inflammatory responses in monocytes. Wild-type apoptotic fibroblasts were incubated for 72 h until most lactate dehydrogenase was released in the extracellular medium. The post-apoptotic cell remnants did not promote a strong inflammatory response in monocytes (Fig. 4b); however, fibroblasts treated with TSA while undergoing apoptosis generated secondarily necrotic cell remnants that promoted inflammation as vigorously as primary necrotic cells that were killed by freeze-thawing.

We could not test whether *Hmgbl*^{-/-} mice have a reduced inflammatory response after tissue necrosis, because these mice survive for only a few hours after birth¹³. To provide evidence in an

animal model for the significance of HMGB1 release, we induced massive hepatocyte necrosis in wild-type mice and measured the inflammatory response. Both in humans and rodents, an overdose of the analgesic acetaminophen (AAP; also known as paracetamol) produces large areas of liver necrosis, concomitant with local inflammation, Kupffer cell activation, and the recruitment and sequestration of neutrophils and macrophages into the injured tissue^{18,19}. Liver damage and neutrophil sequestration are strictly proportional until most hepatocytes become necrotic between 12

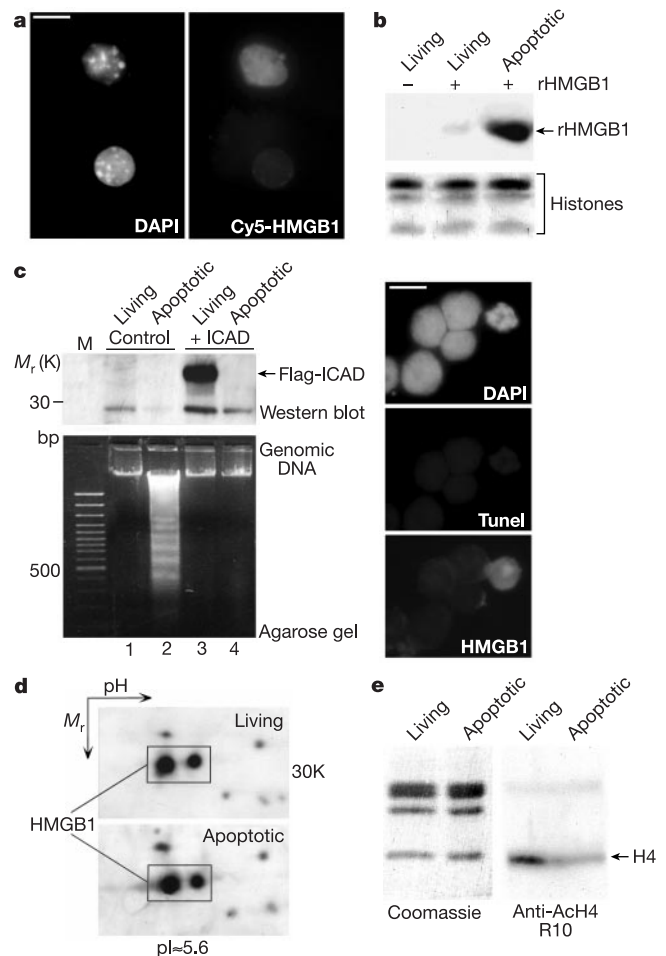


Figure 3 Chromatin changes that occur in apoptosis create binding substrates for HMGB1. Scale bar, 9.5 μ m. **a**, **b**, Bacterially produced HMGB1, labelled with Cy5 (**a**) or unlabelled (**b**), binds to the chromatin of apoptotic *Hmgbl*^{-/-} fibroblasts, but not to that of non-apoptotic fibroblasts, as visualized by microscopy (**a**) or western blotting (**b**). Histones were visualized by Coomassie blue staining. **c**, Left, DNA fragmentation is not responsible for HMGB1 binding to apoptotic nuclei. HeLa cells (expressing a tagged form of ICAD, or control) were either induced into apoptosis ('apoptotic', lanes 2 and 4) or mock-treated ('living', lanes 1 and 3). In apoptosis ICAD is cleaved by caspases and loses the Flag tag (western blot, antibody to Flag, lane 4). Agarose gel electrophoresis shows the internucleosomal cleavage of chromosomal DNA in apoptotic wild-type cells (lane 2), and its inhibition in apoptotic ICAD-expressing cells (lane 4). Right, ICAD-expressing apoptotic cells were permeabilized, fixed and stained for DNA, HMGB1 and TUNEL. Whereas all cells are TUNEL-negative, HMGB1 was firmly retained in the nucleus of the cell showing chromatin condensation (top right). **d**, Total extracts from about 5 million living and apoptotic HeLa cells were subjected to two-dimensional electrophoresis and immunoblotted with antibody to HMGB1. Several acetylated forms of HMGB1 are visible, but no difference is detectable between the two samples. **e**, Histone H4 in apoptotic cells is hypoacetylated. Immunoblotting was done with R10 antibody (specific for the acetylated forms of H4).

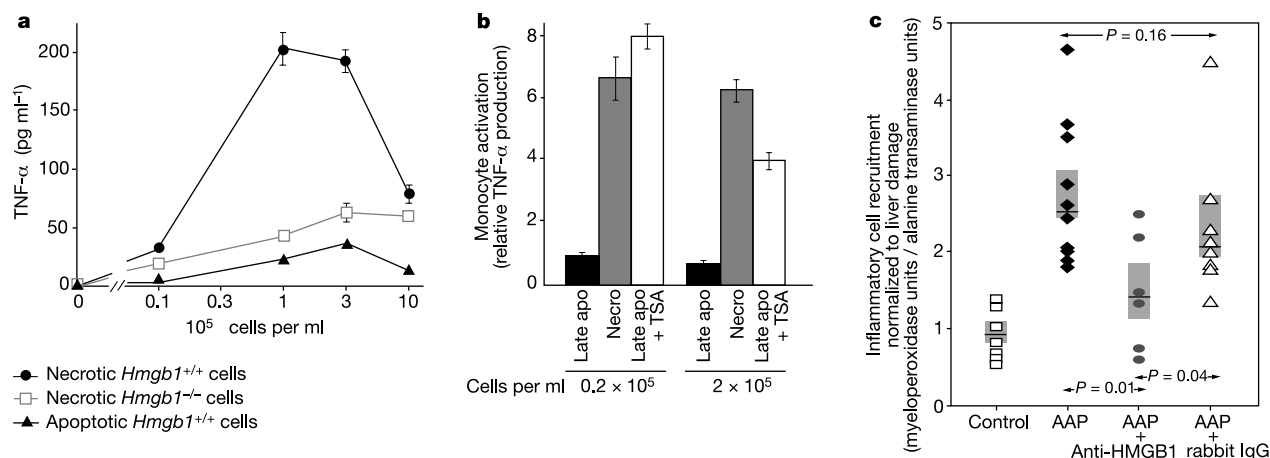


Figure 4 HMGB1 release promotes inflammatory responses. **a**, Necrotic cells lacking HMGB1 do not elicit the production of the pro-inflammatory TNF- α cytokine by monocytes. Bars represent the s.e.m. ($n = 3$). **b**, Apoptotic cells undergoing secondary necrosis and partial autolysis do not promote inflammatory responses unless HMGB1 is mobilized by treatment with TSA. The experiment was repeated three times in duplicate with two different amounts of apoptotic cells to ensure linearity in TNF- α production. Values are normalized to a value of 1 for the amount of TNF- α after challenge with

0.2 $\times 10^5$ apoptotic cells. **c**, Antibodies to HMGB1 reduce inflammation in liver injured by AAP overdose. Liver injury (alanine transaminase activity in serum) and inflammatory cell recruitment (myeloperoxidase activity in total liver extracts) was assessed after 9 h. MPO/ALT ratios indicate inflammation normalized to liver damage. Each point represents one mouse, the bar indicates the median value, and the grey shading indicates the area included in the mean $\pm$ s.e.m. Pairwise comparisons (Mann-Whitney test) between the groups of mice are indicated by the arrows.

and 24 h after AAP poisoning¹⁹. We administered 300 mg per kg (body weight) AAP with a single intraperitoneal injection to young mice, and estimated liver injury after 9 h by measuring alanine transaminase (ALT) activity in serum, and inflammatory cell sequestration by measuring myeloperoxidase (MPO) activity in total liver extracts. Mice received no AAP ($n = 8$), AAP alone ($n = 10$), AAP plus affinity-purified antibodies to HMGB1 (300 mg per kg; $n = 6$), or AAP plus irrelevant rabbit antibodies (300 mg per kg; $n = 8$). All three groups injected with AAP had higher amounts of ALT than sham-treated controls, but the differences between the three treated groups were not statistically significant. Thus, antibodies do not protect against liver damage, at least at the onset of the inflammatory response.

We used the MPO/ALT ratio to compare inflammatory cell recruitment normalized to the amount of liver damage (Fig. 4c). Antibodies to HMGB1 were effective in reducing inflammation after AAP-induced liver necrosis: these mice showed a significantly reduced MPO/ALT ratio (1.5 ± 0.3) in comparison to both mice injected with AAP alone (2.7 ± 0.3 , $P < 0.05$) and mice injected with AAP and pre-immune rabbit IgGs (2.4 ± 0.3 , $P < 0.05$). No HMGB1 could be derived from activated monocytes and macrophages in our experiment, because HMGB1 secretion from inflammatory cells requires at least 16 h^{4,6}. Thus, HMGB1 acts as an immediate trigger of inflammation, as well as a late mediator of inflammation⁴.

In summary, we have shown that the passive release of an abundant chromatin component can serve as a diffusible signal of unprogrammed death, which can be used as a cue to nearby cells. Core histones, although more abundant, would probably not be good signals of necrosis, as they remain anchored to the insoluble chromatin of necrotic cells. Apoptotic cells are not the result of a present and immediate danger and do not trigger inflammation in physiological conditions. They retain nuclear components until cleared by macrophages or nearby cells that act as semiprofessional phagocytes, which they attract and activate by showing 'eat me' signals²⁰. But apoptotic cells that escape prompt clearance undergo secondary necrosis leading to an increased amount of nuclear autoantibodies²¹, and have been also proposed to have an important pathogenetic role in autoimmune diseases such as lupus²². Thus, the

retention of HMGB1 by apoptotic cells undergoing secondary necrosis represents an additional safeguard against confusing necrotic and apoptotic cells. □

Methods

Nomenclature

High mobility group proteins have been renamed²³. High mobility group 1 protein is now officially designated as HMGB1; alternative names are HMG1, amphoterin and p30. HMG-14 and HMG-17 are now HMGN1 and HMGN2.

Constructs and cells

We generated pEGFP-HMGB1 by inserting the coding sequence of the cDNA for rat HMGB1 into pEGFP-N1 (Clontech) using the *EcoRI* and *SacI* restriction sites. Plasmids pEGFP-H1c, pEGFP-NF1, pEGFP-HMGN2 and pEF-flag-mICAD were kindly provided by A. Gunjan, N. Bhattacharyya, R. Hock, M. Bustin and S. Nagata^{11,12,16}. HeLa cells and fibroblasts (line VA1, *Hmgb1*^{+/+}, and line C1, *Hmgb1*^{-/-}) were grown as described^{7,13}. We electroporated HeLa cells with pEGFP-HMGB1 and observed them after 18 h. The average amount of HMGB1-GFP in the cell population was between 1 and 3% of HMGB1 (by immunoblotting with antibody to HMGB1). The amount of HMGB1-GFP varied at most tenfold between different cells; care was taken to use cells with a moderate amount of fluorescence for analysis.

We induced apoptosis by treating the cells for 16 h with 2 ng ml⁻¹ human TNF- α and 35 μ M cycloheximide. Necrosis was induced by treatment for 16 h either with 5 μ M ionomycin and 20 μ M carbonyl cyanide 3-chlorophenylhydrazone (CCCP), or 6 mM deoxyglucose and 10 mM sodium azide, or by three cycles of freezing and thawing.

Three different clones of HeLa cells stably transfected with pEF-flag-mICAD were analysed by TdT-mediated dUTP nick end labelling (TUNEL; Apoptosis Detection System, Promega), and their chromosomal DNA was extracted and separated by electrophoresis on a 1.5% agarose gel.

Indirect immunofluorescence was done as described⁷ using a polyclonal antibody to HMGB1 (Pharmingen) at 1:1,600 dilution, and fluorescein isothiocyanate (FITC)- or tetramethylrhodamine isothiocyanate (TRITC)-conjugated antibodies against rabbit IgG (Boehringer) at 1:300 dilution.

In vivo microscopy and FLIP

Cells were plated and observed in LabTek II chambers (Nalgene) with an Axiovert 135M microscope (Zeiss). We carried out FLIP experiments out on a Leica TCS-SP confocal microscope using the 488-nm excitation line of an Ar laser and detection at 500–575 nm as described¹¹. Cells were bleached (in a spot with radius 1 μ m, 20 mW nominal output, 200–500 ms) and imaged (at 0.2 mW nominal output) at intervals of 6 s.

Binding of recombinant HMGB1 to chromatin

We treated *Hmgb1*^{-/-} fibroblasts with 2 ng ml⁻¹ hTNF- α and 35 μ M cycloheximide. After 16 h, apoptotic cells were recovered by gentle flushing of the dish. Ten million apoptotic *Hmgb1*^{-/-} fibroblasts and a control population of non-apoptotic ones were resuspended in 50 μ l of PBS containing 0.32 M sucrose, 0.5% NP-40 and 1 μ M bacterially produced HMGB1, either fluorescently labelled with Cy5 (Pharmacia) or unlabelled. Average

labelling was 2.3 Cy5 molecules per HMGB1 molecule. After 30 min at room temperature, sample cells were mixed and mounted on slides using Vectashield (Vector Laboratories) containing $1.5 \mu\text{g ml}^{-1}$ 4',6-diamidino-2-phenylindole dihydrochloride (DAPI), and observed on an Axiophot microscope with a TRITC filter (Carl Zeiss). The two pools of cells incubated with unlabelled HMGB1 were layered onto discontinuous gradients formed by 5 ml of 1.16 M sucrose in PBS and a 6-ml cushion of 2 M sucrose in PBS, and centrifuged at 30,000g for 90 min in a SW27 Beckman rotor. We recovered apoptotic and non-apoptotic chromatin that was free of membrane debris from the bottom of the tubes and applied it to a 12% SDS-PAGE gel. The amount of recombinant HMGB1 bound to apoptotic and non-apoptotic chromatin was determined by immunoblotting using an antibody to HMGB1 (Pharmingen) at 1:3,000 dilution. Aliquots of apoptotic and non-apoptotic chromatin were also probed with antibodies to acetyl-histone H4 (R10, a gift from B. Turner), to acetyl-histone H3 (Lys 9, Biolabs) and to acetyl-lysine (Biolabs).

Inflammation assays

To measure TNF- α production *in vitro*, bone marrow was recovered from the hind legs of female C56Bl6 mice, diluted to 5×10^6 cells per ml in Optimum and dispensed in 96-well microtitre plates (120 μl per well). Necrotic cells (lysed by three cycles of freeze-thawing) or apoptotic cells were added to the indicated final concentration into the wells and incubated at 37 °C for 18 h. We assayed TNF- α in the supernatant by enzyme-linked immunosorbent assay (Quantikine M, R&D Systems). TSA was added at 200 ng ml $^{-1}$ together with TNF- α where indicated, and was washed away before mixing the apoptotic cells with bone marrow cells.

To measure inflammation *in vivo*, 1-day-old mice (weighing 1.1 ± 0.1 g) were injected intraperitoneally with 20 μl of PBS containing 320 μg of acetaminophen (Sigma) and 320 μg of antibodies (Pharmingen BD) where indicated. After 9 h, the mice were analysed for serum ALT activity with the GP-Transaminase kit (Sigma) and for MPO activity in liver extracts as described²⁴. We used the non-parametric Mann-Whitney test for statistical analysis of MPO/ALT ratios. Similar results were obtained using Student's *t*-test on the MPO amounts of mice that were paired to minimize the difference in ALT amounts.

Received 12 March; accepted 9 April 2002; doi:10.1038/nature00858.

1. Bustin, M. Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. *Mol. Cell. Biol.* **19**, 5237–5246 (1999).
2. Bianchi, M. E. & Beltrame, M. Upwardly mobile proteins. The role of HMG proteins in chromatin structure, gene expression and neoplasia. *EMBO Rep.* **1**, 109–114 (2000).
3. Hori, O. *et al.* The receptor for advanced glycation end products (RAGE) is a cellular binding site for amphoterin. *J. Biol. Chem.* **270**, 25752–25761 (1995).
4. Wang, H. *et al.* HMG-1 as a late mediator of endotoxin lethality in mice. *Science* **285**, 248–251 (1999).
5. Abraham, E., Arcaroli, J., Carmody, A., Wang, H. & Tracey, K. J. HMG-1 as a mediator of acute lung inflammation. *J. Immunol.* **165**, 2950–2954 (2000).
6. Andersson, U. *et al.* High Mobility Group 1 protein (HMG-1) stimulates proinflammatory cytokine synthesis in human monocytes. *J. Exp. Med.* **192**, 565–570 (2000).
7. Degryse, B. *et al.* The high mobility group (HMG) boxes of the nuclear protein HMG1 induce chemotaxis and cytoskeleton reorganization in rat smooth muscle cells. *J. Cell Biol.* **152**, 1197–2006 (2001).
8. Müller, S. *et al.* The double life of HMGB1 chromatin protein: architectural factor and extracellular signal. *EMBO J.* **16**, 4337–4340 (2001).
9. Falcioni, L. *et al.* High mobility group 1 (HMG1) protein is not stably associated with the chromosomes of somatic cells. *J. Cell Biol.* **137**, 19–26 (1997).
10. Zappavigna, V., Falcioni, L., Helmer Citterich, M., Mavilio, F. & Bianchi, M. E. HMG1 cooperates with HOX proteins in DNA binding and transcriptional activation. *EMBO J.* **15**, 4981–4991 (1996).
11. Phair, R. D. & Misteli, T. High mobility of proteins in the mammalian cell nucleus. *Nature* **404**, 604–609 (2000).
12. Misteli, T., Gunjan, A., Hock, R., Bustin, M. & Brown, D. T. Dynamic binding of histone H1 to chromatin in living cells. *Nature* **408**, 877–881 (2000).
13. Calogero, S. *et al.* The lack of chromosomal protein HMG1 does not disrupt cell growth, but causes lethal hypoglycaemia in newborn mice. *Nature Genet.* **22**, 276–280 (1999).
14. Ura, K., Nightingale, K. & Wolffe, A. P. Differential association of HMG1 and linker histones B4 and H1 with dinucleosomal DNA: structural transitions and transcriptional repression. *EMBO J.* **15**, 4959–4969 (1996).
15. Nightingale, K., Dimitrov, S., Reeves, R. & Wolffe, A. P. Evidence for a shared structural role for HMG1 and linker histones B4 and H1 in organizing chromatin. *EMBO J.* **15**, 548–561 (1996).
16. Enari, M. *et al.* A caspase-activated DNase that degrades DNA during apoptosis, and its inhibitor ICAD. *Nature* **391**, 43–50 (1998).
17. Fadok, V. A., Bratton, D. L., Guthrie, L. & Henson, P. M. Differential effects of apoptotic versus lysed cells on macrophage production of cytokines: role of proteases. *J. Immunol.* **166**, 6847–6854 (2001).
18. Thomas, S. H. L. Paracetamol (acetaminophen) poisoning. *Pharmacol. Ther.* **60**, 91–120 (1993).
19. Lawson, J. A., Farhood, A., Hopper, R. D., Bajt, M. L. & Jaeschke, H. The hepatic inflammatory response after acetaminophen overdose: role of neutrophils. *Toxicol. Sci.* **54**, 509–516 (2000).
20. Ren, Y. & Savill, J. Apoptosis: the importance of being eaten. *Cell Death Differ.* **5**, 563–568 (1998).
21. Scott, R. S. *et al.* Phagocytosis and clearance of apoptotic cells is mediated by MER. *Nature* **211**, 201–211 (2001).
22. Herrmann, M. *et al.* Impaired phagocytosis of apoptotic cell material by monocyte-derived macrophages from patients with systemic lupus erythematosus. *Arthritis Rheum.* **41**, 1241–1250 (1998).
23. Bustin, M. Revised nomenclature for high mobility group (HMG) chromosomal proteins. *Trends Biochem. Sci.* **26**, 152–153 (2001).
24. Kato, A., Yoshidome, H., Edwards, M. J. & Lentsch, A. B. Regulation of liver inflammatory injury by signal transducer and activator of transcription-6. *Am. J. Pathol.* **157**, 297–302 (2000).

Acknowledgements

We thank I. Benzon and L. Ronfani for help with mouse experiments; T. Bonaldi, A. Gunjan, N. Bhattacharyya, R. Hock, M. Bustin, S. Nagata, F. Curnis, A. Corti and B. M. Turner for reagents; V. Caiola for help with preparing Cy5–HMGB1 and statistical analysis; S. Müller for instruction in preparing mouse bone marrow; and D. Donato for personal and technical support. A. Vecchi and A. Mantovani contributed unpublished results; and P. Dellabona, A. Manfredi, R. Pardi and several members of our group provided invaluable suggestions. This work was supported by the Italian Association for Cancer Research and by the Ministry for Education, University and Research.

Competing interests statement

The authors declare competing financial interests: details accompany the paper on *Nature's* website (<http://www.nature.com/nature>).

Correspondence and requests for materials should be addressed to M.E.B. (e-mail: bianchi.marco@hsr.it).

Reciprocal regulation of CD4/CD8 expression by SWI/SNF-like BAF complexes

Tian H. Chi*, Mimi Wan*, Keji Zhao†, Ichiro Taniuchi‡, Lei Chen*, Dan R. Littman‡ & Gerald R. Crabtree*

* Department of Pathology and Developmental Biology, Howard Hughes Medical Institute, Stanford University Medical School, Palo Alto, California 94305, USA

† Laboratory of Molecular Immunology, NHLBI, NIH, Bethesda, Maryland 20892, USA

‡ Molecular Pathogenesis Program, Skirball Institute of Biomolecular Medicine and Howard Hughes Medical Institute, New York University School of Medicine, New York, New York 10016, USA

Thymic development produces two sub-lineages of T cells expressing either CD4 or CD8 co-receptors that assist antibody production and mediate cell killing, respectively. The mechanisms for mutually exclusive co-receptor expression remain poorly defined^{1,2}. We find that mutations in the high mobility group (HMG) domain of BAF57—a DNA-binding subunit of the mammalian SWI/SNF-like chromatin-remodelling BAF complexes—or in the BAF complex ATPase subunit Brg, impair both CD4 silencing and CD8 activation. Brg is haploinsufficient for CD8 activation, but not for CD4 silencing, whereas BAF57 mutations preferentially impair CD4 silencing, pointing to target- and subunit-specific mechanisms of chromatin remodelling. BAF complexes directly bind the CD4 silencer, but the BAF57 HMG domain is dispensable for tethering BAF complexes to the CD4 silencer or other chromatin loci *in vivo*, or for remodelling reconstituted templates *in vitro*^{3,4}, suggesting that chromatin remodelling *in vivo* requires HMG-dependent DNA bending. These results indicate that BAF complexes contribute to lineage bifurcation by reciprocally regulating lineage-specific genes, reminiscent of the role of the yeast SWI/SNF complex in mediating mating-type switching^{5,6}.

As germline deletion of several subunits of BAF complexes lead to either early embryonic lethality^{7–9} or produce little phenotype¹⁰, we studied the developmental roles of the complexes in the T-cell lineage by inactivating the HMG protein BAF57. The Lck proximal promoter was used to direct expression in transgenic mice of dominant negative mutants of BAF57 lacking the entire amino terminus (BAF57 Δ N) or bearing a point mutation (K112I) that disrupts DNA binding³ (Fig. 1a). Expression of either mutant suppressed endogenous BAF57 expression by up to tenfold apparently through autoregulation (Fig. 1b), indicating that the

BAF57 ΔN transgene potentially inactivates 90% of BAF complexes. Immunoprecipitation of BAF complexes from thymic nuclear extracts using the J1 anti-BRG antibody revealed that the *BAF57* ΔN protein was assembled into BAF complexes, because the abundance of *BAF57* ΔN relative to the endogenous *BAF57* and BRG in the immunopurified complexes (bound) was comparable to that in the crude nuclear extract (Fig. 1c, compare lane 2 with 4). The immunopurified BAF complexes were stable: when challenged with 2 M NaCl, the endogenous subunits remained stoichiometrically bound to BRG, although most of the mutant *BAF57* protein dissociated from the complexes (Fig. 1c, lanes 5 and 6). We conclude that T-cell-specific expression of *BAF57* mutants gave rise to complexes specifically deficient in HMG-mediated functions.

Adult *BAF57* transgenic mice showed only a mild (less than twofold) reduction in total thymic cellularity, but the CD4 and CD8 expression pattern was significantly altered, which was best revealed in T-cell antigen receptor (TCR) $\beta^{-/lo}$ cells (Fig. 2, upper panel). Specifically, in *BAF57* transgenic mice, there appeared a TCR $\beta^{-/lo}$ CD4⁺ CD8⁻ population that merged into the CD4⁺ CD8⁺ (double positive) population, the latter showing reduced CD8 but enhanced CD4 expression (compare columns 1 and 3). These changes were synthetically enhanced by a *Brg* null allele⁷ (compare columns 3 and 4); *Brg*^{+/-} mice were viable, but also contained TCR $\beta^{-/lo}$ CD4⁺ CD8⁻ cells whose abundance was similar to that in *BAF57* ΔN heterozygous mice (*BAF57* ΔN ^{+/-}; not shown). The synthetic effect indicates that *BAF57* mutations specifically impaired the function of BAF complexes. Altered CD4 and CD8 expression was not a consequence of global developmental defects, because there was little change in the expression of CD25 (Fig. 3a), TCR β , CD69 (Supplementary Information Fig. 1), CD3, HSA, interleukin (IL)-7R or Bcl2 (not shown). However, mature TCR⁺ thymocytes were significantly reduced in number (up to 30-fold, bottom panel), and CD8 T cells showed a mild and variable increase in CD4 expression (middle panel).

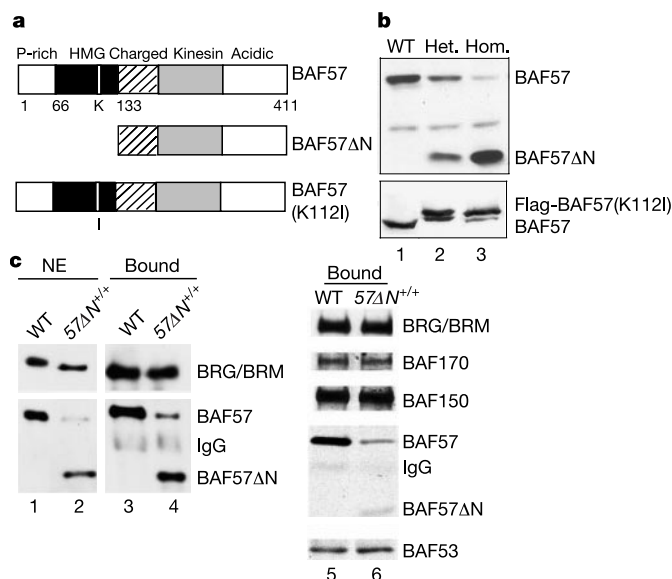


Figure 1 The *BAF57* transgenes suppressed endogenous *BAF57* expression and produce BAF complexes deficient in HMG functions. **a**, The *BAF57* deletion and point mutants. **b**, Western blot analysing *BAF57* expression in thymocytes from wild-type (WT), heterozygous (Het.) or homozygous (Hom.) *BAF57* ΔN (top panel) or *BAF57*(K112) (bottom panel) transgenic mice. The transgenes were Flag-tagged. **c**, Western blot analysing the BAF complexes in thymic nuclear extracts (NE) from wild-type or homozygous *BAF57* ΔN (*57* ΔN ^{+/+}) mice before (lanes 1–2) or after (lane 3–6) immunoprecipitation with the J1 anti-BRG antibody. The beads were washed with RIPA buffer containing either 0.3 M NaCl (lanes 3–4) or 2 M NaCl (lanes 5–6). IgG, immunoglobulin G.

Thymic development begins with CD4⁻ CD8⁻ double negative (DN) T cells that can be subdivided into four stages (DN1–4) (see Methods). Rearrangement of the *TCR* β gene at the DN3 stage leads to CD8 expression (at the CD4⁻ CD8⁺ immature single positive stage) and subsequently CD4 derepression (at the double positive stage) before the double positive cells undergo positive selection to become single positive cells with mutually exclusive co-receptor expression. We thus analysed CD4 and CD8 expression at each of these stages and found that BAF mutations compromised both CD4 silencing and CD8 expression (see below).

The defect in CD4 silencing was first evident as premature CD4 expression at the DN3 stage, with 19% and 49% of DN3 cells expressing CD4 in *BAF57* ΔN ^{+/-} and *BAF57* ΔN homozygous (*BAF57* ΔN ^{+/+}) mice, respectively (Fig. 3a, column 2). No defect was observed in DN1 or DN2 cells, apparently because of lower transgene expression in these cells (not shown). Indeed, CD4 was derepressed in DN1–2 cells when both copies of the *Brg* gene were deleted from these cells using Cre-mediated excision (T.H.C. *et al.*, manuscript in preparation). The *Brg*^{+/-} allele, which did not by itself derepress CD4 (Fig. 3c), enhanced the CD4 level on the *BAF57* ΔN ^{+/-} background such that CD4 was expressed in 25% of DN3 cells in *BAF57* ΔN ^{+/-}; *Brg*^{+/-} mice (Fig. 3a, column 2). CD4 was further derepressed in the mutants as the cells progressed beyond the DN3 stage, as indicated by disappearance of the DN4 cells in the post-DN3 pool. The DN4 cells presumably expressed CD4 prematurely and were thus converted to the CD3⁻ CD4⁺ CD8⁻ cells in the BAF mutants (column 3). However, the CD3⁻ CD4⁺ CD8⁻ cells were far more abundant than the DN4 cells,

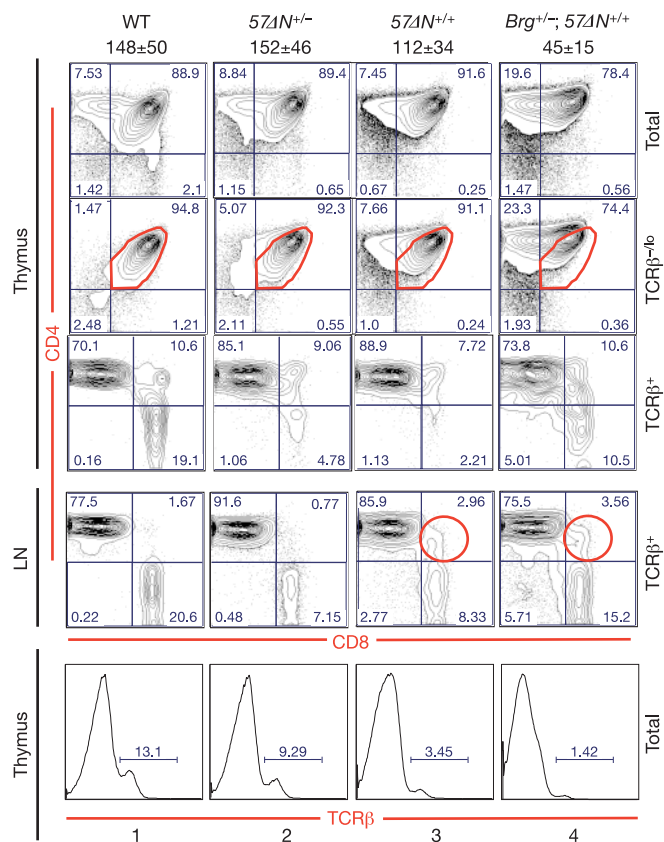


Figure 2 BAF complexes are required for thymic development. Co-receptor (upper and middle panels) and TCR β (bottom panel) expression of various mutant mice. Genotypes of the mice and the total thymocyte numbers are indicated on top of each column. Cells were stained with antibodies against CD4, CD8 and TCR β . The red contours highlight reduced CD8 and increased CD4 expression, whereas the red circles denote the CD4⁺ CD8⁺ lymph node (LN) T cells.

constituting up to 53% of the post-DN3 blast population and 23% of total immature thymocytes in *BAF57ΔN^{+/+}; Brg^{+/-}* mice (Figs 3a (column 3) and 2 (column 4)), suggesting an additional defect in BAF mutants (see below). As expected, CD4 was also expressed in DN3 cells from the *BAF57ΔN^{+/+}; Brg^{+/-}; Rag2^{-/-}* mice, where thymocytes did not develop beyond the double negative stage (Fig. 3b).

In contrast to enhanced CD4 expression, CD8 expression was impaired in BAF mutants. This was first suggested by the fact that the double positive cells from BAF mutants expressed lower levels of CD8 and merged into the novel CD3⁻ CD4⁺ CD8⁻ population, which appeared to be produced by a further decline in CD8 expression to below detectable levels (Figs 2 and 3a). If these changes reflect a specific defect in CD8 expression, one would predict that

they might be synthetically aggravated in combination with loss of one CD8α allele¹¹. This was indeed the case (Fig. 3e), confirming that both the reduction in CD8 levels in double positive cells and the concomitant accumulation of CD3⁻ CD4⁺ CD8⁻ cells originated from a CD8 defect, and suggesting that the CD3⁻ CD4⁺ CD8⁻ cells had accumulated owing to impaired or delayed CD8 expression in developing double positive cells. Indeed, the CD3⁻ CD4⁺ CD8⁻ cells expressed CD8 to become double positive when cultured *in vitro*, but the efficiency was compromised by BAF mutations in a dose-dependent manner (Fig. 3d), consistent with the CD3⁻ CD4⁺ CD8⁻ cells comprising mainly (early) double positive cells lacking detectable CD8 expression. Although CD4 derepression can potentially kill CD8^{hi} double positive cells by inhibiting TCR signalling¹², impaired CD8 expression in double positive cells was unlikely to be

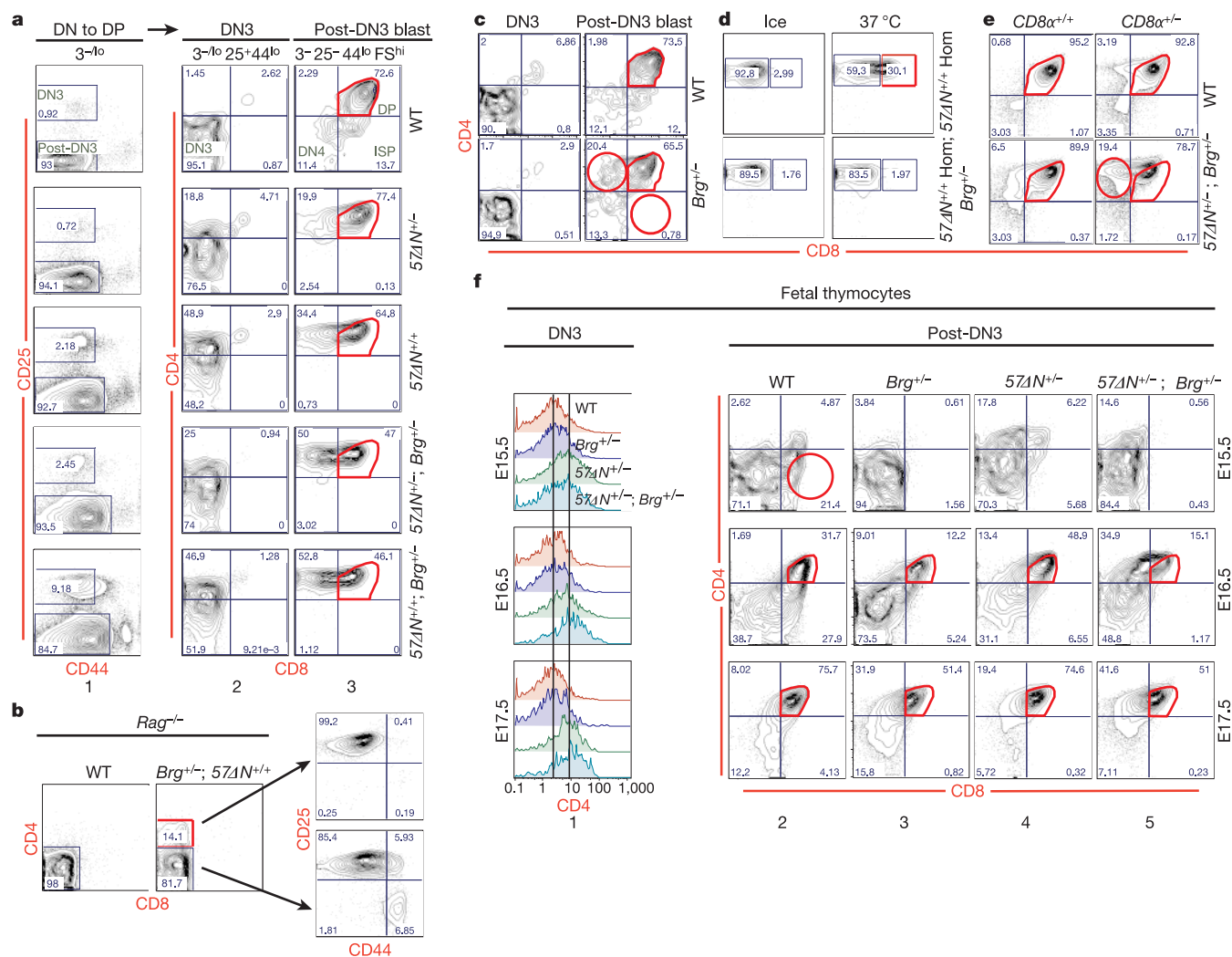


Figure 3 Reciprocal CD4 and CD8 misregulation in the BAF mutants. **a**, CD4 and CD8 expression on DN3 and post-DN3 blasts as defined by CD3 (3), CD25 (25), CD44 (44) and forward scatter (FS) (see Methods for details). DN, double negative; DP, double positive; ISP, immature single positive. **b**, CD4 derepression in *BAF57ΔN^{+/+}; Brg^{+/-}* mice crossed onto the *Rag2^{-/-}* background. The CD4⁺ and CD4⁻ thymocytes in the BAF mutant were further analysed for CD25 versus CD44 expression, which revealed that the CD4⁺ cells were exclusively at the DN3 stage (right panel). **c**, The *Brg^{+/-}* allele selectively suppressed CD8 without derepressing CD4. The circles denote the disappearance and appearance of immature single positive and TCR⁻ CD4⁺ CD8⁻ cells, respectively, and the contours indicate reduced CD8 expression on double positive cells. **d**, *In vitro* differentiation of the TCR⁻ CD4⁺ CD8⁻ cells. Sorted thymocytes from mutant mice were cultured at 37 °C or kept on ice overnight, restained with CD4 and CD8

antibodies and analysed. **e**, Genetic interaction between the BAF complexes and the CD8α gene. *BAF57ΔN^{+/+}; Brg^{+/-}* and *CD8α^{-/-}* mice were crossed to each other and to wild-type mice to produce mice with the indicated genotypes. The cells were stained with CD4, CD8α and TCRβ antibodies, and CD4 and CD8 expression on immature T cells were compared. The circle denotes synthetic increase in the abundance of TCR⁻ CD4⁺ CD8⁻ cells in the BAF, CD8α double mutant, and the contours highlight the synthetic CD8 expression defect in double positive cells. **f**, Co-receptor expression during fetal development. *BAF57ΔN^{+/+}; Brg^{+/-}* males were crossed to wild-type female mice that were killed at various times as indicated. Fetal thymi were assessed as in Fig. 3a except that the post-DN3 cells included the entire CD3^{lo} population, as opposed to the CD3⁻ FS^{hi} blastic subset in adult mice. The circle denotes immature single positive cells.

an artefact resulting from death of CD8^{hi} double positive cells, because there was no significant defect in the viability of double positive cells in BAF mutants (not shown), and CD4 misregulation does not influence CD8 expression^{13–16} (see also Fig. 4, column 2). Consistent with these data, the heterozygous *Brg* deletion did not impair CD4 silencing (Fig. 3c) and yet produced the CD3⁺ CD4⁺ CD8⁺ population while eliminating the CD8^{hi} double positive cells as effectively as the *BAF57ΔN* heterozygous allele (Fig. 3a). These data indicate that BAF mutations impair CD8 expression at the double positive stage independently of CD4 derepression.

CD8 expression was even more susceptible to BAF mutations at the earlier immature single positive (CD3⁺ CD4⁺ CD8⁺) stage, as demonstrated by the complete and paradoxical loss of immature single positive cells—a critical developmental intermediate—in every BAF mutant examined (Fig. 3a, column 3). The loss of immature single positive cells is not due to premature CD4 expression in these CD4⁺ CD8⁺ cells, which could have turned them into 'double positive' cells, because the heterozygous *Brg* deletion completely eliminated the immature single positive cells without derepressing CD4 (Fig. 3c), whereas a CD4 silencer mutation (I.T. and D.R.L., manuscript in preparation) partially depressed CD4 without eliminating the immature single positive cells (Fig. 4a, column 2). Indeed, the triad of concurrent thymic developmental defects—that is, the reduction in CD8 expression in the double positive cells, emergence of the CD3⁺ CD4⁺ CD8⁺ population and loss of immature single positive cells—were each exquisitely sensitive to BAF mutations and were observed in *Brg*^{+/-} mice lacking any other detectable phenotypes including CD4 derepression. Together, these defects demonstrate that the BAF mutations specifically impaired CD8 expression at both immature single positive and double positive stages, and that *Brg* is haploinsufficient for CD8 activation. By these criteria, *Brg* is a major regulator of CD8 expression.

A prediction from the analysis of co-receptor expression in adult mice is that BAF mutations should alter the timing of CD4 and CD8 expression during fetal development. In wild-type controls (Fig. 3f, column 2), CD4 became detectable at embryonic day (E)16.5 as double positive cells emerged, whereas immature single positive cells expressing CD8 (red circle) had already appeared by E15.5, comprising 20% of the total thymic population. Notably, in *BAF57ΔN*^{+/-} and *BAF57ΔN*^{+/-}; *Brg*^{+/-} littermates, CD4 was prematurely derepressed at both DN3 (column 1) and post-DN3 (column 4–5) stages, the effect reaching a plateau by E15.5 in DN3 cells. Immature single positive cells, however, were virtually missing throughout development in *Brg*^{+/-} and *BAF57ΔN*^{+/-}; *Brg*^{+/-} mice, which consequently lacked CD8 expression until E16.5 when CD8^{int/lo} double positive cells emerged (columns 3 and 5). These data demonstrate premature CD4 but delayed CD8 expression in BAF mutants. A second prediction is that the CD3⁺ CD4⁺ CD8⁺ cells, representing double positive cells lacking CD8, should develop with comparable kinetics as the CD4⁺ CD8⁺ double positive cells. Indeed, in *Brg*^{+/-} mice where the CD3⁺ CD4⁺ CD8⁺ pool was free of DN4/immature single positive cells prematurely expressing CD4, these cells emerged at E16.5 and further accumulated at E17.5, concurrent with double positive cells (column 3). Note that although heterozygous *Brg* deletion effectively impaired CD8 expression without derepressing CD4 (column 3), the *BAF57ΔN* transgene preferentially derepressed CD4 while only mildly repressing CD8 during fetal development (column 4). The differential requirement of BAF subunits in gene regulation suggests that chromatin remodelling at distinct loci involves different mechanisms.

In contrast to significantly altered CD4 and CD8 expression during thymic development, co-receptor expression in mature T cells was only slightly affected (Fig. 2, middle panel), apparently because BAF mutations blocked the double positive to single

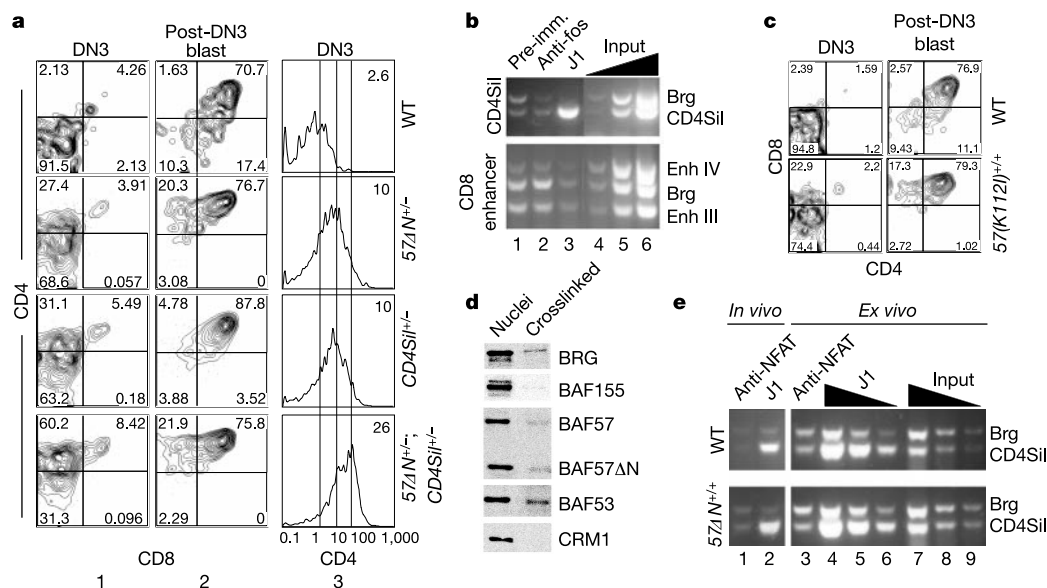


Figure 4 BAF complexes bind the CD4 silencer independently of the BAF57 HMG domain. **a**, Synthetic CD4 derepression in mice carrying mutations in both BAF complexes and the CD4 silencer (CD4Sil). The histograms quantify the mean fluorescence levels of the DN3 populations. **b**, BAF complexes bound the CD4 silencer but not the CD8 enhancers in the M1200 thymoma line²³. Shown are multiplex PCRs measuring the abundance of CD4 silencer (top) or CD8 enhancers (bottom) in crosslinked chromatin before (input) or after immunoprecipitation with various antibodies. A region in the *Brg* genomic sequence was co-amplified to control for DNA yields. The amplification was in the linear range of PCR as shown by a threefold titration of immunoprecipitated input (lanes 4–6). **c**, The *BAF57(K112)* point mutation produced the same defects as the *BAF57* deletion mutant.

d, Western blot analysing the BAF subunits in the nuclei or crosslinked chromatin from *BAF57ΔN*^{+/-} mice. The crosslinking was specific, as CRM1, an abundant nuclear protein, was not detected. The amount of the nuclei loaded on the gel was equivalent to one-tenth of the chromatin sample in terms of DNA quantity. **e**, Chromatin immunoprecipitation showing that the BAF57 HMG domain is dispensable for binding the CD4 silencer. Thymocytes were crosslinked either by perfusing the anaesthetized mice with 2% formaldehyde (*in vivo*) or as single cell suspensions (*ex vivo*) before immunoprecipitation. The amplification was in the linear range of PCR as shown by a threefold titration of immunoprecipitated input (lanes 8 and 9) and precipitated chromatin (lanes 4–6).

positive transition (Fig. 2, bottom panel), thus providing selective pressure for a mature T-cell population dominated by relatively normal T cells.

The CD4 gene is actively suppressed by a 434-base pair silencer^{1,2,13}. To determine whether BAF complexes controlled CD4 expression through the CD4 silencer, we crossed *BAF57ΔN^{+/-}* mice with mice bearing point mutations in the endogenous CD4 silencer (I.T. and D.R.L., manuscript in preparation). The CD4 silencer and *BAF57ΔN^{+/-}* mutations each partially derepressed CD4 on DN3 cells (Fig. 4a, column 1), leading to a 3.8-fold enhancement in the mean fluorescence intensity in CD4 expression (from 2.6 U in the wild type to 10 U in the mutants, column 3), but in combination, they produced a tenfold increase (from 2.6 to 26 U). This suggests a genetic interaction between BAF complexes and the CD4 silencer, and raises the possibility that BAF complexes directly bind the CD4 silencer. Indeed, J1 anti-BRG or anti-BAF57 antibodies, but not irrelevant antibodies (pre-immune, anti-Fos or anti-NFATc2), selectively precipitated the CD4 silencer in a thymoma line (Fig. 4b, upper panel; data not shown). This binding was specific, because neither CD8 enhancer III nor IV (ref. 1) bound BAF complexes (Fig. 4b, lower panel).

The BAF57 deletion mutant lacked the entire N terminus including the HMG and the proline-rich domains. Mice transgenic for the point mutant BAF57(K112I) also showed reciprocal CD4 and CD8 misregulation (Figs 1a and 4c), demonstrating a specific requirement for HMG function *in vivo*. The HMG domain is dispensable for tethering BAF complexes to chromatin in general, because BAF57ΔN was crosslinked to chromatin as effectively as the endogenous BAF57 (Fig. 4d). It was also dispensable for targeting BAF complexes to the CD4 silencer (Fig. 4e). These results are consistent with the presence of multiple DNA/chromatin binding domains in BAF complexes^{17,18}, and suggest that it is the DNA bending activity unique to the HMG domain that is required for chromatin remodelling *in vivo*.

Studies of the developmental roles of the SWI/SNF-like complexes in multicellular organisms have been hampered by the early lethality in *Drosophila*¹⁹ and mammals⁷⁻⁹. Our results reveal that a conserved function of yeast and mammalian complexes is to influence lineage differentiation. Furthermore, the fact that BAF complexes can reciprocally regulate co-receptor expression provides a new entry point for studying the mechanisms of positive selection, as BAF complexes are known to respond to TCR stimulation in peripheral T cells²⁰. The fact that the Brg homologue Brm is dispensable for thymic development (M. Yaniv and M. Malissen, personal communication) reinforces the idea that combinatorial assembly of BAF complexes gives rise to divergent biological functions. □

Methods

Mice

BAF57 transgenic mice were produced by standard methods and maintained on the CD1 background.

Flow cytometry

For co-receptor expression in adult mice, single-cell suspensions were prepared and typically subjected to five-colour analysis using anti-CD4-APC, anti-CD8-Texas red, Anti-CD25-fluorescein isothiocyanate (FITC), anti-CD44-phycoerythrin (PE), anti-CD3-Cy5, anti-B220-Cy5 and propidium iodide. Normal thymic development before the single positive stage proceeds from CD3⁺ CD25⁺ CD44⁺ (DN1) to CD3⁺ CD25⁺ CD44⁺ (DN2) to CD3⁺ CD25⁺ CD44^{low} (DN3) to CD3⁺/low CD25⁺ CD44^{low} (post-DN3); the post-DN3 cells comprise DN4, immature single positive (CD3⁺ CD4⁺ CD8⁺) and double positive cells¹. We resolved immature thymocytes—marked by no to low CD3 expression and thus including all the cells from the double negative to double positive stage—into DN1–3 and post-DN3 subsets (Fig. 3a, column 1) based on CD25 and CD44 expression, and we determined CD4 and CD8 expression in each of these subsets (Fig. 3a, columns 2 and 3). For post-DN3 cells, we analysed post-DN3 blasts, the most immature cells in the post-DN3 pool that are marked by high forward scatter (FS) and no CD3 expression; this was necessary to enrich for and reveal the rare DN4 and immature single positive cells. Note the elevation in CD44 expression in post-DN3 cells and correspondingly enlarged post-DN3 gates in the mutants (Fig. 3a, column 1); this adjustment did not alter flow

cytometric profiles of post-DN3 cells. Co-receptor expression during fetal development was analysed exactly as in adult mice, except that the gate for post-DN3 cells was set on the entire CD3⁺/low^{int} cells, as opposed to the blastic subset in this population.

Chromatin immunoprecipitation and analysis of crosslinked proteins

Chromatin immunoprecipitation (ChIP) was performed as described²², except that the crosslinked chromatin was purified using a CsCl gradient, the chromatin was precleared with protein A pre-immune antibody complex, and that BSA and salmon sperm DNA were added during immunoprecipitation. To analyse the proteins in the chromatin fraction purified with the CsCl₂ gradient, a portion of the chromatin was precipitated with acetone, the pellet resuspended in 2 × SDS-gel loading buffer and incubated at 65 °C overnight to reverse crosslinking and solubilize the proteins before analysis.

Received 4 April; accepted 21 May 2002; doi:10.1038/nature00876.

1. Ellmeier, W., Sawada, S. & Littman, D. R. The regulation of CD4 and CD8 coreceptor gene expression during T cell development. *Annu. Rev. Immunol.* **17**, 523–554 (1999).
2. Sarafova, S. D. & Siu, G. Control of CD4 gene expression: connecting signals to outcomes in T cell development. *Braz. J. Med. Biol. Res.* **32**, 785–803 (1999).
3. Wang, W. *et al.* Architectural DNA binding by a high-mobility-group/kinesin-like subunit in mammalian SWI/SNF-related complexes. *Proc. Natl Acad. Sci. USA* **95**, 492–498 (1998).
4. Phelan, M. L., Sif, S., Narlikar, G. J. & Kingston, R. E. Reconstitution of a core chromatin remodeling complex from SWI/SNF subunits. *Mol. Cell* **3**, 247–253 (1999).
5. Stern, M., Jensen, R. & Herskowitz, I. Five SWI genes are required for expression of the HO gene in yeast. *J. Mol. Biol.* **178**, 853–868 (1984).
6. Sudarsanam, P. & Winston, F. The Swi/Snf family nucleosome-remodeling complexes and transcriptional control. *Trends Genet.* **16**, 345–351 (2000).
7. Bultman, S. *et al.* A Brg1 null mutation in the mouse reveals functional differences among mammalian SWI/SNF complexes. *Mol. Cell* **6**, 1287–1295 (2000).
8. Knochendler-Yeivin, A. *et al.* The murine SNF5/INI1 chromatin remodeling factor is essential for embryonic development and tumour suppression. *EMBO Rep.* **1**, 500–506 (2000).
9. Roberts, C. W., Galusha, S. A., McMenamin, M. E., Fletcher, C. D. & Orkin, S. H. Haploinsufficiency of Snf5 (integrator interactor 1) predisposes to malignant rhabdoid tumors in mice. *Proc. Natl Acad. Sci. USA* **97**, 13796–13800 (2000).
10. Reyes, J. C. *et al.* Altered control of cellular proliferation in the absence of mammalian brahma (SNF2α). *EMBO J.* **17**, 6979–6991 (1998).
11. Fung-Leung, W. P. *et al.* CD8 is needed for development of cytotoxic T cells but not helper T cells. *Cell* **65**, 443–449 (1991).
12. Nakayama, T. *et al.* Inhibition of T cell receptor expression and function in immature CD4⁺CD8⁺ cells by CD4. *Science* **249**, 1558–1561 (1990).
13. Zou, Y. R. *et al.* Epigenetic silencing of CD4 in T cells committed to the cytotoxic lineage. *Nature Genet.* **29**, 332–336 (2001).
14. Leung, R. K. *et al.* Deletion of the CD4 silencer element supports a stochastic mechanism of thymocyte lineage commitment. *Nature Immunol.* **2**, 1167–1173 (2001).
15. Teh, H.-S. *et al.* Participation of CD4 coreceptor molecules in T-cell repertoire selection. *Nature* **349**, 241–243 (1991).
16. Davis, C. B., Killeen, N., Crooks, M. E., Raulet, D. & Littman, D. R. Evidence for a stochastic mechanism in the differentiation of mature subsets of T lymphocytes. *Cell* **73**, 237–247 (1993).
17. Agaloti, T. *et al.* Ordered recruitment of chromatin modifying and general transcription factors to the IFN-β promoter. *Cell* **103**, 667–678 (2000).
18. Nie, Z. *et al.* A specificity and targeting subunit of a human SWI/SNF family-related chromatin-remodeling complex. *Mol. Cell Biol.* **20**, 8879–8888 (2000).
19. Tamkun, J. W. *et al.* brahma: a regulator of *Drosophila* homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. *Cell* **68**, 561–572 (1992).
20. Zhao, K. *et al.* Rapid and phosphoinositide-dependent binding of the SWI/SNF-like BAF complex to chromatin after T lymphocyte receptor signalling. *Cell* **95**, 625–636 (1998).
21. Godfrey, D. I., Kennedy, J., Suda, T. & Zlotnik, A. A developmental pathway involving four phenotypically and functionally distinct subsets of CD3⁺ CD4⁺ CD8⁺ triple-negative adult mouse thymocytes defined by CD44 and CD25 expression. *J. Immunol.* **150**, 4244–4252 (1993).
22. Liu, R. *et al.* Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. *Cell* **106**, 309–318 (2001).
23. Sawada, S., Scarborough, J. D., Killeen, N. & Littman, D. R. A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development. *Cell* **77**, 917–929 (1994).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank C. Wilson for encouragement and comments; K. Akashi for help with FACS during the early course of the work; K. Stankunas, J. Owen, A. Singer, I. Graef, K. Neu and C. Forsberg for critical readings of the manuscript; T. Magnuson for Brg null mice before publication; R. Perlmutter for Lck promoter constructs; and H. Liu for technical help. T.C. is supported by a fellowship from Leukemia and Lymphoma Society of America. G.R.C. is a HHMI investigator. These studies were funded by a grant from the National Institute of Health.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.R.C. (e-mail: crabtree@cmgm.stanford.edu).

Apoptosis disables CD31-mediated cell detachment from phagocytes promoting binding and engulfment

Simon Brown*, Isabelle Heinisch*, Ewan Ross†, Kate Shaw*, Chris D. Buckley† & John Savill*

* Inflammation Repair Group, MRC Centre for Inflammation Research, University of Edinburgh, Edinburgh EH8 9AG, UK

† Division of Immunity & Infection, MRC Centre for Immune Regulation, University of Birmingham, Birmingham B15 2TT, UK

Macrophage recognition and ingestion of 'self' cells undergoing apoptosis *in vivo* protects tissues from the toxic contents of dying cells and modulates macrophage regulation of inflammatory and immune responses^{1,2}. However, the complex molecular mechanisms mediating macrophage discrimination between viable and apoptotic cells are poorly understood^{2,3}. In particular, little is known of why viable nucleated cells are not engulfed by macrophages. To reveal active repulsion of viable cells and to seek specific capture or 'tethering' of apoptotic cells, we studied macrophage binding of viable and apoptotic leukocytes under conditions of flow. We found that homophilic ligation of CD31 (ref. 4) on viable leukocytes promoted their active, temperature-dependent detachment under low shear, whereas such CD31-mediated detachment was disabled in apoptotic leukocytes, promoting tight binding and macrophage ingestion of dying

cells. Here we propose that CD31 (also known as platelet-endothelial cell adhesion molecule-1, PECAM-1) is an example of a cell-surface molecule that prevents phagocyte ingestion of closely apposed viable cells by transmitting 'detachment' signals, and which changes function on apoptosis, promoting tethering of dying cells to phagocytes.

We adapted a flow chamber system⁵ to study human leukocyte binding to human macrophage monolayers (Fig. 1). At low shear (32 mPa) apoptotic but not viable neutrophils (polymorphonuclear neutrophils (PMNs)) bound to THP-1 macrophage monolayers at 37 °C (Fig. 1a). However, at 20 °C (chosen to inhibit cytoskeletal reorganization) both viable and apoptotic PMNs bound macrophages equally well at different loading shear (Fig. 1b). This suggested that whereas viable and apoptotic leukocytes used similar binding mechanisms at 20 °C, viable leukocytes were selectively capable of active, temperature-dependent detachment from macrophages that would have been overlooked in conventional 'static' phagocytic assays in which viable cells are not ingested⁶. To test this idea, we bound viable human leukocytes (PMNs or Jurkat T lymphocytes) to monolayers of macrophages (THP-1 or primary human monocyte-derived macrophages (HMDMs)) at 20 °C under low shear (Fig. 1c–e). When temperature in the cell chamber was raised to 37 °C, while maintaining constant shear, viable but not apoptotic leukocytes detached. This was not a nonspecific effect of raising temperature because identically loaded PMNs did not detach from adhesive glass coverslips mounted in the chamber (data not shown).

To elucidate which surface molecules were involved in mediating tethering at low shear, we used five-day-old HMDMs at 20 °C to screen for apoptotic PMN proteins that mediated stable binding of plasma-membrane-enriched proteoliposomes. By differentially extracting macrophage-bound proteins with detergents we purified a single major polypeptide that remained insoluble to Triton-X100. This was identified as CD31 by retarded electrophoretic mobility

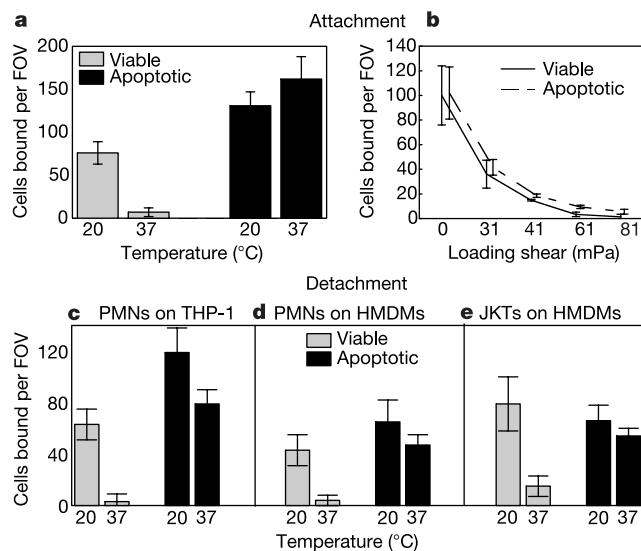


Figure 1 Specific failure of viable leukocytes to attach to macrophage monolayers under flow conditions at 37 °C reflects active, temperature-dependent detachment of viable but not apoptotic leukocytes. **a**, In attachment assays, viable, but not apoptotic, PMNs exhibited minimal binding to THP-1 macrophages at 37 °C despite binding at 20 °C. **b**, Binding of viable or apoptotic PMNs to THP-1 macrophages at 20 °C was performed at various loading shear, and results normalized for a count of 100 per field of view (FOV) at 0 mPa, being 76 ± 13 and 131 ± 16 , respectively. **c**, **d**, In detachment assays, viable or apoptotic PMNs were passed over a monolayer of THP-1 macrophages (**c**) or five-day-old HMDMs (**d**) in 10% FCS/DMEM at 32 mPa, and bound cells counted before and after raising the temperature. **e**, Alternatively, viable and apoptotic Jurkat cells were passed over HMDM monolayers. All error bars are s.d. for $n = 3$ separate PMN, THP-1 and HMDM preparations; differences between viable leukocyte binding at 37 °C and 20 °C were statistically significant at $P < 0.001$ (Tukey's Honestly Significant Difference (HSD) test).

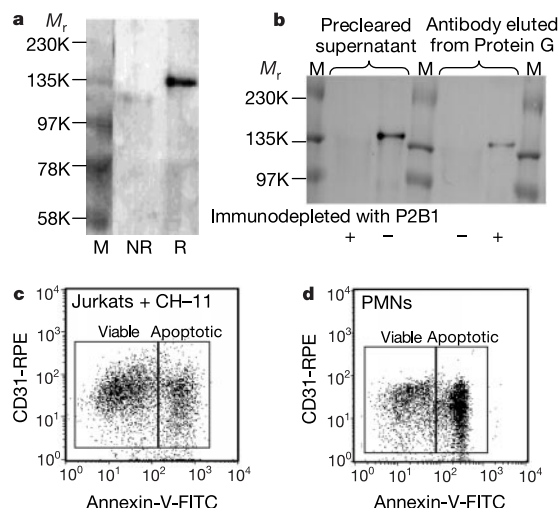


Figure 2 CD31 mediates tethering of proteoliposomes derived from apoptotic PMNs by macrophages. **a**, Primary human monocyte-derived macrophages bound a major apoptotic neutrophil-derived polypeptide that migrated with an atypical increase in size under reducing (R) conditions compared with non-reducing conditions (NR), suggestive of CD31 (ref. 21); molecular markers are shown at M. **b**, A CD31-specific monoclonal antibody, P2B1, immunodepleted the partially purified protein recovered from Protein G with 0.1 M glycine, pH 3.0. **c**, **d**, Levels of CD31 expression by Jurkat T cells (**c**) and PMNs (**d**) were not altered significantly after apoptosis, as assessed by flow cytometry using an R-phycoerythrin-conjugated (RPE) anti-CD31 monoclonal antibody, WM-59 (CD31-RPE). In the experiment shown, mean peak fluorescence for viable and apoptotic cells was 53 and 25 (Jurkat T cells) and 54 and 20 (PMNs).

upon reduction, and by specific immunodepletion and recovery (Fig. 2a, b). Notably, although there was small-scale shedding of CD31 from apoptotic leukocytes consistent with membrane loss by blebbing (Fig. 2c), immunoprecipitation revealed no evidence of CD31 cleavage during leukocyte apoptosis (not shown).

CD31 is a member of the immunoglobulin superfamily of transmembrane cell-surface adhesion and signalling molecules that is expressed by both leukocytes and macrophages⁴. Homophilic interaction of leukocyte CD31 with endothelial CD31 regulates leukocyte motility and active movement across endothelial cell surfaces away from the initial site of binding, directing leukocytes towards intercellular junctions^{7,8}. Therefore, we tested whether homophilic CD31 interactions might mediate both leukocyte binding and selective, active detachment of viable but not apoptotic leukocytes from a bound state. When loaded and washed at low shear and at 20 °C in 'attachment' assays (Fig. 3a), both viable and apoptotic leukocytes specifically bound to non-adhesive Thermanox coverslips coated with human recombinant CD31 (but not to uncoated (shown) or intracellular adhesion molecule-1 (ICAM1)-coated Thermanox coverslips (data not shown)). Moreover, at 37 °C only apoptotic leukocytes bound CD31 significantly. In 'active detachment' assays (Fig. 3b, c), viable and apoptotic leukocytes both attached to CD31 at constant low shear and at 20 °C, but increasing the temperature to 37 °C caused detachment of viable leukocytes only. Importantly, viable leukocytes did not detach from fibronectin-coated Thermanox coverslips at 37 °C (not shown).

The inferred role for leukocyte CD31 in both attachment and detachment from macrophage CD31 was confirmed by study of 'CD31-negative' Jurkat lines (serially selected for minimum CD31 expression⁹); these failed to bind to CD31-coated Thermanox coverslips at low shear at 20 °C unless transfected with wild-type CD31, in which case they bound with similar efficiency to parent (CD31-positive) Jurkat cells (Fig. 3c). Furthermore, such transfectants not only detached from CD31 at 37 °C when viable, but also failed to detach at 37 °C when induced to undergo apoptosis. Crucially, when transfected with mutant glycosyl phosphatidylinositol (GPI)-linked CD31 constructs, viable CD31-negative Jurkat cells bound to CD31 at low shear at 20 °C but did not detach at 37 °C (Fig. 3c). In addition, viable Jurkat cells transfected with GPI-

linked CD31 also failed to detach from macrophage monolayers (37.8 ± 1.6 at 20 °C, 49.8 ± 5.5 at 37 °C, mean $\pm$ s.d.) unlike wild-type CD31 transfectants (40.0 ± 8.9 at 20 °C, 11.0 ± 5.1 at 37 °C). GPI-linked CD31 is a good model for 'signalling-disabled' CD31 as, compared with wild-type CD31, it is expressed at similar levels and distribution, retains normal homophilic binding, but fails to signal, for example, Rap1 and integrin activation^{9,10}. Indeed, a double tyrosine to phenylalanine mutant of the ITIM motif within the cytoplasmic tail of CD31 (Y663F Y686F), which also fails to signal Rap1 and integrin activation^{9,10}, also failed to confer detachment from macrophage monolayers at 37 °C (37.8 ± 14.2) when expressed in CD31-negative Jurkat cells.

These data indicated that active, temperature-dependent leukocyte detachment from CD31 required normally functioning CD31 and implied that CD31-mediated signalling might be disrupted in apoptotic leukocytes. In support of the latter, immunoprecipitation demonstrated that CD31 expressed by apoptotic (but not viable) leukocytes failed to associate constitutively with the cytoplasmic signalling molecule SHP-1, and failed to bind SHP-2 on pervanadate treatment (Fig. 4)¹¹. Furthermore, agonistic antibody ligation of CD31 on apoptotic leukocytes also specifically failed to induce calcium transients (not shown). Nevertheless, further work will be required to understand how apoptosis can disable CD31-signalled cell detachment from homophilic binding with CD31, allowing for a 'switch' in CD31 function.

Having demonstrated the discriminatory properties of attachment/detachment mediated by apoptotic compared with viable leukocyte CD31 interaction with macrophage CD31 under conditions of flow, we sought evidence that this interaction contributed to macrophage ingestion of apoptotic leukocytes in a more conventional 'static' assay in which viable leukocytes were not ingested (not shown). Apoptotic CD31-negative Jurkat cells were ingested, but with lesser efficiency than when transfected with wild-type or GPI-linked CD31, and with lesser efficiency than parent (CD31-positive) Jurkat cells—blocking antibody to CD31 reduced ingestion to the level seen with CD31-negative apoptotic cells in all cases (Fig. 5a). Closely comparable data were obtained in macrophage ingestion assays of apoptotic PMNs (Fig. 5b, c), in which ingestion was inhibited by pretreatment of either CD31 partner with blocking antibodies. This is consistent with homophilic CD31 interaction adding to the efficiency of ingestion in a system in which static assays have defined phagocyte roles for macrophage vitronectin receptor and CD36 (refs 6, 12, 13).

We show that CD31 mediates phagocyte 'tethering' of apoptotic cells under conditions of flow, which at many sites *in vivo* will be important in allowing phagocytes subsequently to engage molecular 'eat me' signals displayed by apoptotic cells (such as phosphatidylserine^{2,14}). More importantly, our data also demonstrate that

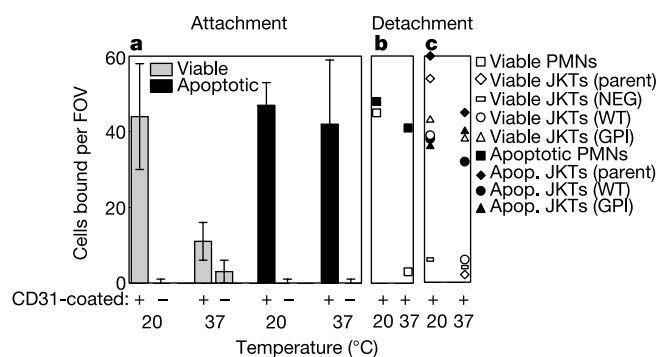


Figure 3 Leukocyte interaction with CD31-coated coverslips under flow demonstrates that homophilic CD31 interactions mediate both attachment of apoptotic cells and detachment of viable cells. **a**, Viable or apoptotic PMNs were passed over Thermanox coverslips that were either sham-treated or coated with CD31. **b, c**, Temperature-dependent detachment of PMNs (**b**) or various Jurkat (JKT) T-cell lines (**c**) bound to CD31 was also assessed. Jurkat lines used were CD31-positive (parent), from which a CD31-negative clone had been selected (NEG) and subsequently stably transfected with wild-type CD31 (WT) or a cytoplasmic tailless GPI-anchored form (GPI)^{9,10}. Note that the latter lacked the capacity to direct temperature-dependent detachment of viable leukocytes; all other viable leukocytes demonstrated statistically significant detachment at 37 °C. Symbols depict mean, *n* was a minimum of three separate experiments, and standard deviation was never greater than 15% of the mean.

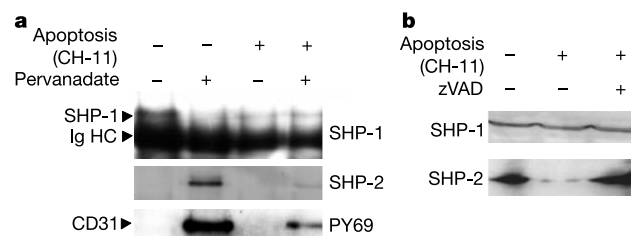


Figure 4 Apoptosis disables CD31-mediated signalling in leukocytes. **a**, Immunoblots demonstrate that apoptotic leukocytes show reduced constitutive association of SHP-1 with CD31 (upper panel), reduced pervanadate-mediated recruitment of SHP-2 to CD31 (middle panel) and reduced CD31 tyrosine phosphorylation (lower panel). Ig HC, immunoglobulin heavy chain. **b**, Total SHP-2 protein, but not SHP-1, was found to decrease in apoptotic Jurkat cells that were inhibited by carbobenzoxy-Val-Ala-Asp-fluoromethylketone (zVAD).

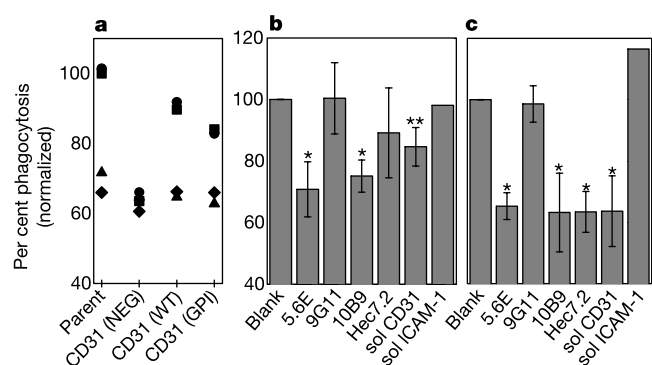


Figure 5 CD31 on apoptotic leukocytes interacts with CD31 on phagocytes to promote engulfment. **a**, Engulfment of apoptotic Jurkat cells by five-day-old HMDMs in the presence of no antibody (squares) or anti-CD31 monoclonal antibodies 9G11 (non-blocking, circles), 10B9 (blocking, diamond), or 5.6E (blocking, triangle), all at $5 \mu\text{g ml}^{-1}$. Data is normalized to apoptotic parent Jurkat ingestion ($34.7 \pm 7.6\%$, mean $\pm$ s.d., $n = 3$), the inhibitory effects being statistically significant ($P < 0.01$, Tukey's HSD test). Standard deviation was never greater than 12% of the mean. **b**, **c**, Flow cytometric assessment of engulfment by HMDMs pretreated (**b**) or not (**c**) with apoptotic PMNs that were pretreated (**b**) or not (**c**) with denoted agents; Hec7.2 is a blocking anti-CD31 monoclonal antibody. Microscopic examination of cytospin preparations confirmed apoptotic PMNs were ingested and at levels with those enumerated by flow cytometry. The level of phagocytosis in the blank controls was $28.6 \pm 4.4\%$ (mean $\pm$ s.d.). All experiments are a minimum of $n = 5$. Asterisk, significantly different from blank control at $P < 0.02$; double asterisk, significantly different from blank control at $P < 0.05$ (Tukey's HSD test). sol CD31 and ICAM-1 are human recombinant soluble proteins.

homophilic CD31 interaction can discriminate between apoptotic and viable cells by selectively imparting 'detachment' signals to viable cells, a mechanism that prevents macrophage ingestion of viable self-cells. This may complement the display by viable cells of molecules that ligate macrophage receptors inhibiting phagocytosis¹⁵. How apoptosis switches CD31 function from 'repulsive' to 'adhesive' will require further study, but may help explain why molecules in other families, such as thrombospondin, can increase viable leukocyte motility¹⁶ while also promoting phagocytosis of apoptotic cells of various lineages^{12,13}, and should prompt a search for analogous systems in cells other than leukocytes. □

Methods

Leukocyte isolation and culture

Human PMNs and monocytes were isolated from freshly drawn venous blood and separated after citration, dextran sedimentation, and discontinuous Percoll gradients as described previously¹⁷. Polymorphonuclear neutrophils were cultured overnight at 37°C in Iscove's modified Dulbecco's medium supplemented with 10% autologous platelet-rich, plasma-derived serum (PRPDS) to allow constitutive apoptosis¹⁷. Primary human monocyte-derived macrophages were obtained by culturing adherent monocytes on plastic or 13 mm Thermanox coverslips (Nalge Nunc International) in Iscove's plus 10% PRPDS for five days^{6,11}. THP-1, a human premyelomonocytic leukaemic cell line, and human leukaemic Jurkat T-cell lines, were maintained in 10% FBS/RPMI-1640. Cells from the THP-1 line were differentiated to macrophages with 12-*O*-tetradecanoylphorbol-13-acetate (TPA) (10 ng ml^{-1} , 72 h) and cultured on Thermanox coverslips. Jurkat T cells underwent apoptosis after Fas ligation (CH-11, 100 ng ml^{-1} , 24 h) in the absence of serum. Apoptotic PMNs or Jurkat cells were positively sorted with annexin-V-biotin and streptavidin-conjugated paramagnetic beads, and cells recovered in the absence of calcium (greater than 95% purity).

Protein-coated coverslips

Thermanox coverslips, washed with NaHCO_3 (50 mM, pH 8.2), were coated for 2 h at room temperature with $1 \mu\text{g}$ human recombinant soluble CD31 or ICAM-1 (R&D Systems); made up in PBS at 1 mg ml^{-1} and diluted 200-fold with NaHCO_3 (50 mM, pH 8.2). Alternatively, coverslips were coated with $5 \mu\text{g ml}^{-1}$ human plasma fibronectin in PBS (Calbiochem). Coverslips were washed with PBS and blocked with 10% FCS/DMEM for 30 min at room temperature.

Attachment and detachment assays

Macrophage monolayers or coated Thermanox coverslips were placed in a modified flow cell chamber supporting laminar flow⁵ and monitored by video microscopy. Leukocytes

in 10% FCS/DMEM, pre-warmed to either 20°C or 37°C , were delivered by means of a three-way solenoid valve and fluid flow maintained under negative pressure using a syringe-driven pump (Harvard 2000). Typically a 1-ml bolus of 2×10^6 cells was followed by a minimum of 1 ml medium to flush out unbound cells before those cells which remained firmly bound over a further 2-min period were counted. For attachment assays, and unless otherwise stated, cells were introduced to the flow chamber under a steady flow (0.17 ml min^{-1}) conferring 32 mPa shear force. In detachment assays, binding of viable or sorted apoptotic cells was achieved at 20°C and 32 mPa, binding was then checked and counted before the temperature was raised to 37°C by switching the input medium and allowing the flow cell to equilibrate at 32 mPa for a further 5 min before counting. In all cases results are presented as the mean $\pm$ s.d. per field of view with $\times 40$ objective.

Proteoliposomes and identification of CD31 as a tethering protein

Unsorted apoptotic PMNs, surface biotinylated with NHS-LC-biotin ($1 \text{ mg per } 10^8$ cells), were re-suspended after washing in ice-cold lysis buffer (10 mM HEPES, pH 7.4, 1 mM EDTA, 1 mM benzamide, 1 mM σ -phenanthroline, 1 mM phenylmethylsulphonyl-fluoride, $10 \mu\text{M}$ pepstatin, $10 \mu\text{M}$ leupeptin, $10 \mu\text{M}$ antipain) at 20×10^6 cells ml^{-1} (ref. 18). The cell pellet was re-suspended in lysis buffer before sonication and sequential centrifugation at $10,000g$ (10 min) and $500,000g$ (10 min). The microsomal-enriched pellet at $500,000g$ was sonicated and washed twice before re-suspending by sonication in HBSS, before incubating with HMDMs at 20°C and extracting with 1% Triton-X100 followed by RIPA buffer. The RIPA extract was then analysed after SDS-polyacrylamide gel electrophoresis (PAGE) and western blot for biotinylated proteins under non-reducing and reducing conditions. Partial purification of biotinylated proteins was achieved by sequential application to and recovery from discontinuous 1% SeaKem LE/4% MetaPhor agarose tube gels (12 mm internal diameter) according to the supplier's instructions (FMC BioProducts), run under non-reducing and reducing conditions.

Immunoprecipitation and western blotting

CD31 was immunoprecipitated from precleared extracts of equivalent numbers of viable or unsorted apoptotic leukocytes with 9G11 after treatment of intact cells with or without sodium pervanadate and re-suspension in the above lysis buffer containing 1% Triton-X100, 1 mM sodium vanadate, 1 mM sodium fluoride, 1 mM levanisole¹⁹. Immunoprecipitates were washed with PBS before resolving by SDS-PAGE and transfer to polyvinylidene difluoride (PVDF). Blots were then blocked with 2% casein in PBS before probing with anti-SHP-1 (clone 52), anti-SHP-2 (clone79), or anti-phosphotyrosine (PY69) antibodies (all from Transduction Laboratories).

Phagocytosis assays

Polymorphonuclear neutrophils and Jurkat T-cell lines were labelled with ((4-chloromethyl)benzoyl)aminotetramethylrhodamine (Molecular Probes) in serum-free Iscove's or DMEM medium at $2 \mu\text{g}$ of 20×10^6 cells ml^{-1} for 20 min at 37°C (ref. 20), and cultured overnight at 4×10^6 cells ml^{-1} in serum-replete Iscove's or serum-free DMEM and CH-11 (100 ng ml^{-1}), respectively, to promote apoptosis. Apoptotic Jurkat cells, washed but unsorted, were incubated with five-day-old HMDMs for 30 min, unengulfed cells were removed, and adherent HMDMs were collected for flow cytometric analysis^{6,17,20} (Becton-Dickinson FACSscan). For generation of apoptotic PMNs, aged PMNs (washed but not sorted) or five-day-old HMDMs were pre-incubated separately for 10 min at room temperature with antagonistic (5.6E, PECAM1.3, 10B9, Hec7.2) or neutral (9G11) CD31 monoclonal antibodies ($5 \mu\text{g ml}^{-1}$) and recombinant CD31 or ICAM-1 protein ($10 \mu\text{g ml}^{-1}$), before washing and use in phagocytosis assays.

Received 19 November 2001; accepted 22 March 2002; doi:10.1038/nature00811.

- Kerr, J. F., Wyllie, A. H. & Currie, A. R. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br. J. Cancer* **26**, 239–257 (1972).
- Savill, J. & Fadok, V. Corpse clearance defines the meaning of cell death. *Nature* **407**, 784–788 (2000).
- Green, D. R. & Beere, H. M. Apoptosis. Mostly dead. *Nature* **412**, 133–135 (2001).
- Newman, P. J. The biology of PECAM-1. *J. Clin. Invest.* **99**, 3–8 (1997).
- Sakariassen, K. S., Aarts, P. A., de Groot, P. G., Houdijk, W. P. & Sixma, J. J. A perfusion chamber developed to investigate platelet interaction in flowing blood with human vessel wall cells, their extracellular matrix, and purified components. *J. Lab. Clin. Med.* **102**, 522–535 (1983).
- Savill, J., Dransfield, I., Hogg, N. & Haslett, C. Vitronectin receptor-mediated phagocytosis of cells undergoing apoptosis. *Nature* **343**, 170–173 (1990).
- Rainger, G. E., Buckley, C., Simmons, D. L. & Nash, G. B. Cross-talk between cell adhesion molecules regulates the migration velocity of neutrophils. *Curr. Biol.* **7**, 316–325 (1997).
- Muller, W. A., Weigl, S. A., Deng, X. & Phillips, D. M. PECAM-1 is required for transendothelial migration of leukocytes. *J. Exp. Med.* **178**, 449–460 (1993).
- Reedquist, K. A. *et al.* The small GTPase, Rap1, mediates CD31-induced integrin adhesion. *J. Cell Biol.* **20**, 1151–1158 (2000).
- Newton, J. P., Buckley, C. D., Jones, E. Y. & Simmons, D. L. Residues on both faces of the first immunoglobulin fold contribute to homophilic binding sites of PECAM-1/CD31. *J. Biol. Chem.* **272**, 20555–20563 (1997).
- Pumphrey, N. J. *et al.* Differential association of cytoplasmic signalling molecules SHP-1, SHP-2, SHIP and phospholipase C- γ 1 with PECAM-1/CD31. *FEBS Lett.* **450**, 77–83 (1999).
- Savill, J., Hogg, N., Ren, Y. & Haslett, C. Thrombospondin cooperates with CD36 and the vitronectin receptor in macrophage recognition of neutrophils undergoing apoptosis. *J. Clin. Invest.* **90**, 1513–1522 (1992).
- Ren, Y., Silverstein, R. L., Allen, J. & Savill, J. CD36 gene transfer confers capacity for phagocytosis of cells undergoing apoptosis. *J. Exp. Med.* **181**, 1857–1862 (1995).
- Fadok, V. A. *et al.* A receptor for phosphatidylserine-specific clearance of apoptotic cells. *Nature* **405**, 85–90 (2000).

15. Oldenborg, P. A. *et al.* Role of CD47 as a marker of self on red blood cells. *Science* **288**, 2051–2054 (2000).
16. Mansfield, P. J., Boxer, L. A. & Suchard, S. J. Thrombospondin stimulates motility of human neutrophils. *J. Cell Biol.* **111**, 3077–3086 (1990).
17. Savill, J. S. *et al.* C. Macrophage phagocytosis of aging neutrophils in inflammation. Programmed cell death in the neutrophil leads to its recognition by macrophages. *J. Clin. Invest.* **83**, 865–875 (1989).
18. Brown, S. B., Bailey, K. & Savill, J. Actin is cleaved during constitutive apoptosis. *Biochem. J.* **323**, 233–237 (1997).
19. Bird, I. N. *et al.* Homophilic PECAM-1 (CD31) interactions prevent endothelial cell apoptosis but do not support cell spreading or migration. *J. Cell Sci.* **112**, 1989–1997 (1999).
20. Brown, S. B., Clarke, M. C., Magowan, L., Sanderson, H. & Savill, J. Constitutive death of platelets leading to scavenger receptor-mediated phagocytosis. A caspase-independent cell clearance program. *J. Biol. Chem.* **275**, 5987–5996 (2000).
21. Metzelaar, M. J., Korteweg, J., Sixma, J. J. & Nieuwenhuis, H. K. Biochemical characterization of PECAM-1 (CD31 antigen) on human platelets. *Thromb. Haemost.* **66**, 700–707 (1991).

Acknowledgements

We would particularly like to acknowledge C. Stewart, E. Drost and W. MacNee for assisting in establishing a flow-based approach; H. Jersmann for help with a number of FACS-based phagocytosis assays; A. Rossi for help with calcium flux studies; and C. Haslett and M. Salmon for discussion. Work was funded by the Wellcome Trust, the Chief Scientist Office of the Scottish Executive, and a Wellcome Trust International Fellowship to I.H.

Competing interests statement

The authors declare that they have no competing financial interests

Correspondence and requests for materials should be addressed to S.B. (e-mail: simon.brown@ed.ac.uk).

Overexpression of β -carotene hydroxylase enhances stress tolerance in *Arabidopsis*

P. A. Davison, C. N. Hunter & P. Horton

Robert Hill Institute, Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, UK

Plant stress caused by extreme environmental conditions is already a principal reason for yield reduction in crops¹. The threat of global environment change makes it increasingly important to generate crop plants that will withstand such conditions. Stress, particularly stress caused by increased sunlight, leads to the production of reactive oxygen species that cause photo-oxidative cell damage². Carotenoids, which are present in the membranes of all photosynthetic organisms, help protect against such light-dependent oxidative damage³. In plants, the xanthophyll cycle (the reversible interconversion of two carotenoids, violaxanthin and zeaxanthin⁴) has a key photoprotective role⁵ and is therefore a promising target for genetic engineering to enhance stress tolerance. Here we show that in *Arabidopsis thaliana* overexpression of the *chyB* gene that encodes β -carotene hydroxylase—an enzyme in the zeaxanthin biosynthetic pathway⁶—causes a specific twofold increase in the size of the xanthophyll cycle pool. The plants are more tolerant to conditions of high light and high temperature, as shown by reduced leaf necrosis, reduced production of the stress indicator anthocyanin and reduced lipid peroxidation. Stress protection is probably due to the function of zeaxanthin in preventing oxidative damage of membranes.

To investigate the protective role of zeaxanthin under stress conditions, we used metabolic engineering to increase the size of the xanthophyll cycle pool in *A. thaliana*. Given the nutritional importance of carotenoids, several groups have manipulated carotenoid biosynthesis in plants^{7–9}, and overexpression of a bacterial phytoene desaturase gene in tobacco leads to an increase in the

xanthophyll cycle pool¹⁰. Although mutants of *A. thaliana* with an increased xanthophyll cycle pool size are available, they have large alterations in the content of other carotenoids and lesions in the biosynthesis of the plant growth regulator abscisic acid^{2,11}.

To manipulate specifically the size of the xanthophyll cycle pool, we overexpressed the thylakoid membrane enzyme β -carotene hydroxylase, which catalyses the conversion of β -carotene to zeaxanthin (Fig. 1); antisense inhibition of the encoding gene (*chyB*) is already known to result in a reduction in pool size¹². We transformed *A. thaliana* (ecotype C24) with the *chyB* gene under the control of a constitutive promoter (Fig. 2a), established several stable lines derived from independent transformation events (Fig. 2b) and checked these for overexpression of β -carotene hydroxylase (Fig. 2c).

Using the published complementary DNA sequence for β -carotene hydroxylase¹³, we designed primers specific to +418 to +887 in the carboxy-terminal region and carried out polymerase chain reaction with reverse transcription (RT-PCR) using increasing starting amounts of total RNA from wild-type and transformed plants. Densitometry indicated a linear increase in the amount of wild-type β -carotene hydroxylase messenger RNA over the range shown (Fig. 2c), but for the sense *chyB* transformant the amount of detectable mRNA had almost reached saturation point at the lowest amount analysed (0.25 μ g of total RNA), which represented a fourfold increase as compared with wild-type expression.

Lines that were overexpressing the gene were subjected to carotenoid extraction and high performance liquid chromatography (HPLC) to determine their xanthophyll profiles (Table 1). There was a marked (at least twofold) increase in the amount of violaxanthin in the sense *chyB* lines. This would be expected if the conversion of β -carotene to zeaxanthin were enhanced, as the latter would be converted by zeaxanthin epoxidase into violaxanthin (through antheraxanthin) in the dark (Fig. 1). Roughly the same increase was seen in all three lines. This increased pool of viola-

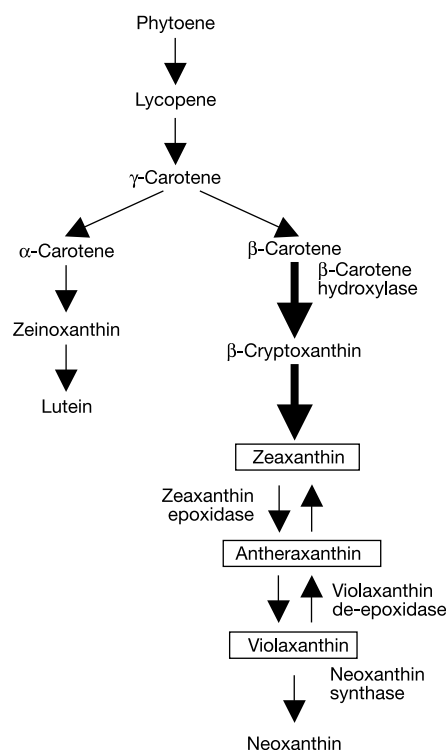


Figure 1 Carotenoid–xanthophyll biosynthetic pathway in plants. The step catalysed by β -carotene hydroxylase is indicated by bold arrows. The xanthophyll cycle carotenoids are shown in boxes.

Table 1 Leaf carotenoid/xanthophyll profiles and photosynthetic parameters of wild-type (C24) and sense *chyB* plants

Line	$\mu\text{g chl } a \text{ per mg dry weight}$	$\text{chl } a/b$	Carotenoid (mmol)/chl <i>a</i> (mol)						Fv/Fm	NPQ
			Viola	Lutein	β -car	Anth	Neo	Total carot		
C24										
Low-light grown	10.5 $\pm$ 1.6	3.05 $\pm$ 0.10	49 $\pm$ 6	166 $\pm$ 8	126 $\pm$ 9	8 $\pm$ 3	46 $\pm$ 4	395	0.81 $\pm$ 0.00	2.33 $\pm$ 0.09
Moderate-light grown			99 $\pm$ 12	199 $\pm$ 3	163 $\pm$ 11	19 $\pm$ 3	66 $\pm$ 2	546		
Sense 1										
Low-light grown	10.0 $\pm$ 0.7	2.97 $\pm$ 0.18	116 $\pm$ 10	164 $\pm$ 6	110 $\pm$ 9	7 $\pm$ 3	48 $\pm$ 3	445	0.80 $\pm$ 0.00	2.25 $\pm$ 0.10
Moderate-light grown			257 $\pm$ 46	180 $\pm$ 4	147 $\pm$ 2	25 $\pm$ 2	65 $\pm$ 3	675		
Sense 2										
Low-light grown	11.1 $\pm$ 1.7	3.19 $\pm$ 0.26	114 $\pm$ 7	154 $\pm$ 12	109 $\pm$ 6	9 $\pm$ 2	44 $\pm$ 4	430	0.80 $\pm$ 0.01	2.05 $\pm$ 0.03
Moderate-light grown			275 $\pm$ 30	194 $\pm$ 6	155 $\pm$ 10	30 $\pm$ 7	66 $\pm$ 4	720		
Sense 3										
Low-light grown	10.0 $\pm$ 2.0	3.00 $\pm$ 0.16	117 $\pm$ 16	160 $\pm$ 11	107 $\pm$ 6	9 $\pm$ 3	50 $\pm$ 6	444	0.81 $\pm$ 0.00	2.48 $\pm$ 0.22
Moderate-light grown			285 $\pm$ 34	186 $\pm$ 18	150 $\pm$ 10	41 $\pm$ 6	65 $\pm$ 6	728		

Values are the means $\pm$ s.d. of at least three 5-week-old, dark-adapted plants grown in low light (100 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$) or in moderate light (400 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$). NPQ was determined after illumination with 6,000 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ actinic light. anth, antheraxanthin; β -car, β -carotene; carot, carotenoid; neo, neoxanthin; viola, violaxanthin. Fv (variable fluorescence/Fm (maximum fluorescence)) is a measure of photosystem II efficiency.

xanthin was not converted further into neoxanthin. Overexpression of β -carotene hydroxylase did not unduly perturb the rest of the carotenoid biosynthetic pathway; there was only a slight fall in the amount of β -carotene, presumably owing to increased conversion to the xanthophylls.

The total carotenoid content per unit of chlorophyll *a* was

increased and, as the chlorophyll content of the plants was unchanged, we concluded that the increase in violaxanthin was not at the expense of other carotenoids. Similar increases in the xanthophyll cycle pool size were obtained regardless of whether the sense *chyB* plants were grown under low (100 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$) or moderate (400 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$) light (Table 1). Hence, for low-light plants the xanthophyll cycle pool increased from 14% to nearly 30% of total carotenoid for low-light plants, and from 22% to over 40% for moderate-light plants. The latter value is significantly larger than any reported value for wild-type *Arabidopsis*.

To determine where this extra xanthophyll was located, we extracted thylakoid membranes, gently solubilized them with detergent, separated them on a sucrose density gradient and carried out HPLC on the resulting fractions (Fig. 3). The extra xanthophyll was mostly associated with the photosystem II light-harvesting complexes (LHCII), which showed a threefold increase in the amount

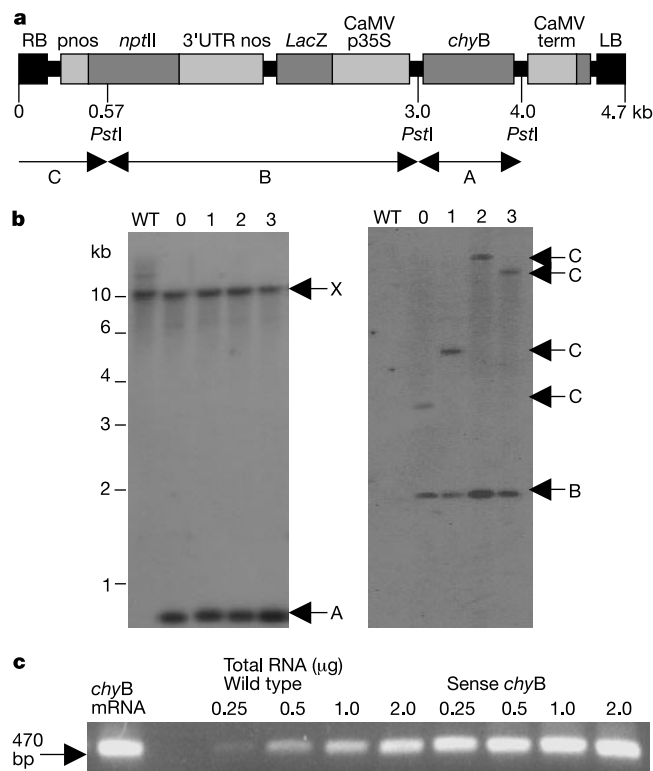


Figure 2 Southern and RT-PCR analysis of 5-week-old sense *chyB* plants grown in low light. **a**, Transfer DNA (TDNA) region of pBin400::*chyB*. LB, left TDNA border; RB, right TDNA border; CaMV p35S, cauliflower mosaic virus 35S promoter; CaMV term, cauliflower mosaic virus 35S terminator; *lacZ*, β -galactosidase gene; *nptII*, kanamycin resistance gene; pnos, nopaline synthase promoter; 3' UTR nos, nopaline synthase terminator. **b**, Southern blot of *Pst*I-cut DNA probed with *chyB* (left) and *nptII* (right), showing the endogenous (X) and introduced (A) *chyB* genes, a *nptII*-containing internal TDNA fragment (B) and a RB *nptII* fragment (C) containing flanking genomic sequence of varying size because of independent insertions into the genome. WT, wild type; 0–3, T_3 sense *chyB* plants descended from separate T_0 transformants. **c**, RT-PCR specific for *chyB* mRNA carried out on increasing amounts of wild-type and sense *chyB* (line 3) total RNA. The PCR products were separated on a 1% agarose gel (equal loadings per lane). A positive control containing about 10 ng of synthesized full-length *chyB* mRNA was included.

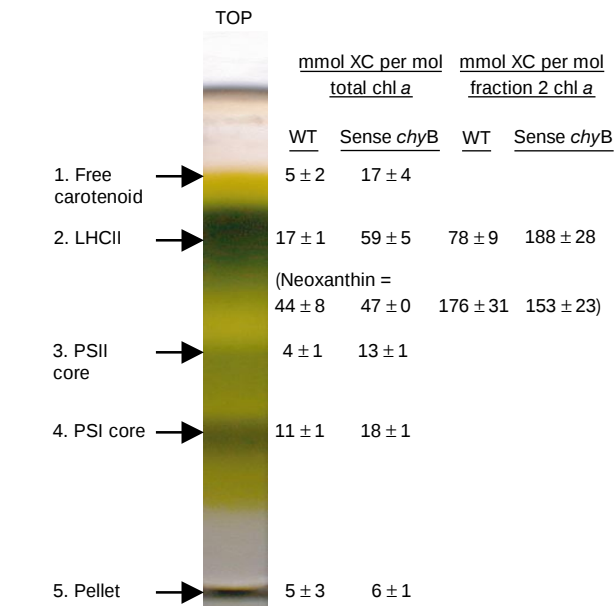


Figure 3 Sucrose density gradient (0.15 M to 1.0 M) profile of thylakoid membranes isolated from dark-adapted 5-week-old, low-light grown wild-type and sense *chyB* plants after solubilization in 1.5% (w/v) dodecyl maltoside. Xanthophyll cycle carotenoids (XC) and neoxanthin (given in parentheses) contents of each fraction are standardized against total moles of chlorophyll *a* (chl *a*) added to the gradient and, where indicated, moles of chl *a* contained in a particular fraction. Values are means $\pm$ s.d. ($n = 4$). PS, photosystem; LHCII, photosystem II light-harvesting complex, which includes both the major and minor complexes.

of bound violaxanthin. Previous work has shown that LHCII contains xanthophyll cycle binding sites at the periphery of the complex that are not fully occupied¹⁴. Thus, the extra violaxanthin associated with LHCII is probably bound to these sites. The ratio of violaxanthin to neoxanthin (1 molecule bound per LHCIIb complex) in the LHCII fraction of the transformed plants (1.2 compared with a wild-type value of 0.4) is consistent with this conclusion.

The extra violaxanthin was available for de-epoxidation when the plants were exposed to high light (15 min, 1,000 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$). The amounts of zeaxanthin were at least twice as large in the sense *chyB* lines as in the wild-type, although their average de-epoxidation state (percentage of total xanthophyll pool present as antheraxanthin and zeaxanthin) was the same (52% and 54%, respectively). We conclude that at least some of the extra xanthophyll must be biologically active, which would allow us to test whether an increased xanthophyll cycle pool will confer any improvement in stress tolerance. One of the ways in which plants respond to stress is to increase the size of their xanthophyll pool and so increase their amount of zeaxanthin^{2,5}. Engineered plants with a constitutively increased pool size should be 'preprimed' and better able to respond to stress conditions. To investigate this, we switched plants grown for 5 weeks under conditions of low light (100 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$) and 20 °C to stress conditions (1,000 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$, air temperature, 40 °C; mean $\pm$ s.d. leaf temperature: wild type, 36.6 $\pm$ 2.4 °C; sense *chyB*, 36.4 $\pm$ 2.5 °C; n = 8) for 2 weeks. As expected, the amount of zeaxanthin was roughly 2–4 times greater in the sense *chyB* lines

than in the wild-type (Fig. 4c).

An indicator of stress in *A. thaliana* is the presence of the purple flavonoid anthocyanin^{15,16}. The flavonoid pathway is distinct from the carotenoid biosynthesis pathway and should not be affected by perturbation of the latter¹⁷. We analysed 24 stressed plants from the control wild type (transformed with the transformation vector alone) and 2 sense *chyB* lines, and determined their relative anthocyanin content per gram of fresh weight (Fig. 4a). There was no difference between the two sense *chyB* populations (mean $\pm$ s.d. 1.0 $\pm$ 0.47 and 0.9 $\pm$ 0.43), but both these populations had significantly less anthocyanin (P < 0.0001, Student's *t*-test) than the wild type (mean 1.6 $\pm$ 0.49). This difference in anthocyanin amounts was obvious in the whole plants (Fig. 4b) where the sense *chyB* lines were clearly greener and healthier, and showed less leaf necrosis.

Zeaxanthin is involved in the non-photochemical quenching (NPQ) mechanism used by plants to dissipate excess light energy^{18,19}. But increased NPQ is unlikely to be the cause of the observed stress tolerance because we detected no increase in the capacity of NPQ in the sense *chyB* plants (Table 1). This is consistent with observations that constitutive zeaxanthin-accumulating *Arabidopsis* mutants do not show an increase in the maximum extent of NPQ^{20,21}. Zeaxanthin also prevents the accumulation of reactive oxygen species that lead to oxidative damage of membranes through lipid peroxidation^{22,23}. When zeaxanthin synthesis is inhibited by mutating violaxanthin de-epoxidase, the increase in oxidative stress under high-light conditions does not arise from the inhibition of NPQ but from the removal of this alternative protective function²². This observation suggested that the increased stress tolerance of the sense *chyB* plants might arise from an enhancement of this protective role of zeaxanthin. The fact that the extra xanthophyll in the sense *chyB* plants seems to occupy the loose peripheral binding sites on LHCII (which are in contact with membrane lipids) is consistent with this suggestion.

As an estimate of general lipid peroxidation, we determined the amount of malondialdehyde (MDA)—a secondary end product of the oxidation of polyunsaturated fatty acids—in wild-type and sense *chyB* plants exposed to stress. We assayed four populations of plants: wild-type, wild-type transformed with pBin400 and two sense *chyB* lines (1 and 3). There was no significant difference in mean MDA content (nmoles per gram of fresh weight; mean $\pm$ s.d.) between the two wild-type populations (wild type, 8.6 $\pm$ 4.7; transformed wild type, 9.7 $\pm$ 3.5) or the two sense *chyB* lines (line 1, 6.5 $\pm$ 4.0; line 3, 6.8 $\pm$ 3.6). But there was a difference in the amount of lipid peroxidation between the wild-type and sense *chyB* lines. Figure 4d shows the pooled wild-type and sense *chyB* MDA content values subdivided into groups; the mean values for the two populations (27 plants per population) were 9.2 $\pm$ 4.3 and 6.6 $\pm$ 3.9, respectively, indicating a significant $\sim$ 30% drop in the amount of lipid peroxidation in the sense *chyB* lines (P < 0.02, Student's *t*-test).

When the production of zeaxanthin is prevented by mutation²² or antisense suppression²⁴ of violaxanthin de-epoxidase, plants have an increased sensitivity to light stress. It has been suggested that zeaxanthin also has a role in protecting thylakoid membranes specifically from the effects of heat stress²⁵. Our results show that increasing expression of the β -carotene hydroxylase enzyme brings about an increase in the content of the xanthophyll cycle and zeaxanthin in the chloroplast membrane, and notably that this manipulation leads to an improved tolerance of high-light and high-temperature conditions. It is important to point out that these effects on anthocyanin production and lipid peroxidation were observed in plants that had been exposed to stress conditions for 2 weeks; that is, it was a sustained property of the transformed plants. We conclude that genetic manipulation of a single enzyme in carotenoid metabolism can bring about a pronounced increase in the stress tolerance of plants and thus represents a potentially powerful way forward in the production of stress-tolerant crops. □

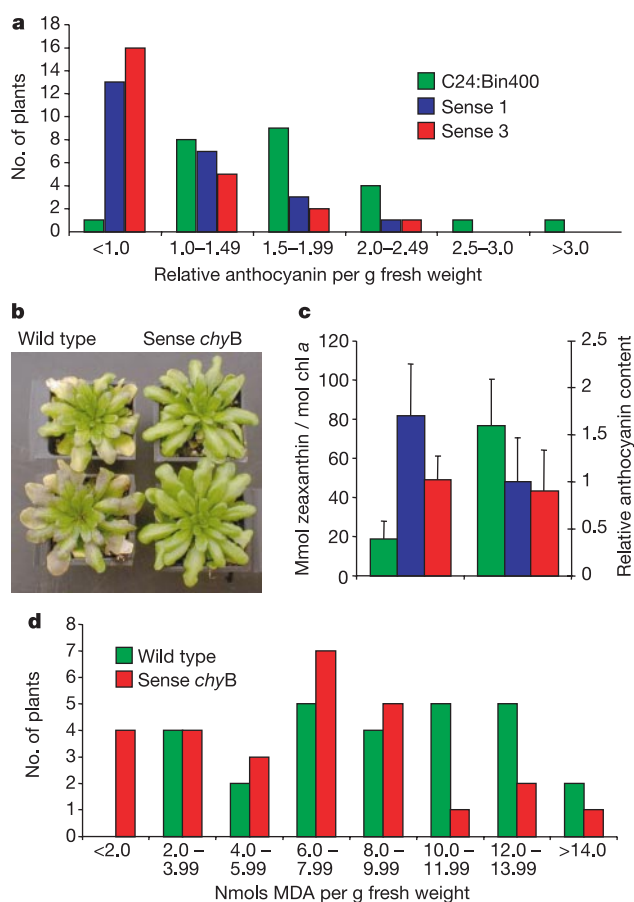


Figure 4 Response of 5-week-old sense *chyB* plants grown in low light subjected to 2 weeks of high-light and high-temperature stress. **a**, Anthocyanin content. C24:Bin400, control wild-type transformed with pBin400. **b**, Physical appearance. **c**, Zeaxanthin content (standardized against chlorophyll *a*) and relative anthocyanin produced per gram of fresh weight. Values are means $\pm$ s.d. (n = 24). **d**, Malondialdehyde (MDA) content per gram of fresh weight.

Methods

Plant material and growth conditions

We grew *A. thaliana* (ecotype C24) on soil:vermiculite:perlite (4:2:1) under a 8 h/16 h light/dark regime at 20 °C and 100 µmol photons m⁻² s⁻¹ (low-light conditions), 20 °C and 400 µmol photons m⁻² s⁻¹ (moderate-light conditions), or 40 °C and 1,000 µmol photons m⁻² s⁻¹ (stress conditions). If selection was required, seed was first germinated on 1 × Murashige–Skooog basal medium supplemented with Gamborgs vitamins, 1% sucrose, 0.5 g l⁻¹ MES and 50 µg ml⁻¹ kanamycin.

Generation of sense *chyB* plants

The *chyB* gene (GenBank accession code U58919) was isolated by PCR (forward primer, 5'-GCCGATCCTGCAGCCTCAAAACAAGTAGCTCCTCTCC-3'; reverse primer, 5'-TTTCTAGACTCGAGTCAAGAACTCGAACTGACCCCGG-3') on a size-selected cDNA library (derived from 1–2 kilobases of mRNA) supplied by the Arabidopsis Biological Resource Center, Ohio State University. The blunt-ended fragment was inserted into the *Sma*I cloning site located in the 35S CaMV expression cassette of pDH51. The 35S CaMV promoter–*chyB*–terminator cassette was isolated using PCR (forward primer, 5'-GCATCGATAAGCTTCAGCTATGACCATGATTACG-3'; reverse primer, 5'-TAAACGACGGCCAGTGCC-3'), confirmed by DNA sequencing and inserted into the *Hind*III site of plant transformation vector pBin400 (supplied by M. Bevan, John Innes Institute). The pBin400::*chyB* construct was transformed into *Agrobacterium tumefaciens* C58C1 (pGV2260; rifampicin resistant) by triparental mating, and transformation of *A. thaliana* was carried out in tissue culture using root explants and shoot regeneration as described²⁶. We selected primary transformants (T₀) on the basis of resistance to the antibiotic kanamycin and established lines by single seed descent, with selection maintained to the third generation (T₃). Twelve T₃ plants per line were analysed by HPLC for presence of the high xanthophyll pool size phenotype to check for homozygosity.

Southern blotting

Genomic DNA was isolated from plants using the GenElute kit (Sigma). We carried out Southern blotting as recommended by Hybond. Radiolabelled probes were produced using the Ready-to-Go kit (Pharmacia).

RT-PCR

Total RNA was extracted from 5-week-old leaf tissue using the RNeasy kit (Qiagen) and treated with RNase-free DNase (Promega). We synthesized control full-length β-carotene hydroxylase mRNA using an RNA transcription kit (Stratagene). First strand cDNA synthesis was carried out using reverse primer 5'-AGTCAAGAACTCGAACTCGACCCGG-3' and avian myeloblastosis virus (AMV) reverse transcriptase as recommended by Promega. Second strand synthesis and amplification was carried out with the reverse primer and forward primer 5'-GGGCTCATAGAGCTCTGTGGC-3' using standard PCR techniques.

Thylakoid membranes, carotenoid extraction and HPLC

We prepared thylakoid membranes as described²⁷. Chlorophyll content was measured spectrophotometrically in 80% acetone²⁸. Carotenoids were extracted from a single leaf in ethanol:ether (1:2 v/v) under red light, filtered through cotton wool and a 0.2-µm syringe filter, dried under nitrogen gas and stored in the dark at -20 °C before being dissolved in acetone. Samples were separated by reverse-phase HPLC on a Spherisorb 5-µm ODS2 column (4.6 × 250 mm, Waters Corp.) using a 40-min gradient of ethyl acetate (0 to 100%) in acetonitrile:water (9:1 v/v) at a flow rate of 1 ml min⁻¹. Pigments were identified by their retention time and by absorption spectra using a photodiode array detector (model 996, Waters Corp.) and were quantified by integrating peak areas. We converted peak areas to molar concentrations by comparison with carotenoid standards of known concentration run on HPLC and standardized against the molar concentration of chlorophyll *a*. To confirm the determinations of zeaxanthin in those cases where it was not fully separated from lutein, two leaf samples were taken from each plant, one under the ambient light conditions and one after a 1-h dark treatment (to convert zeaxanthin to violaxanthin). Any reduction in the xanthophyll pool size in the former compared with that seen in the latter sample was assumed to be due to zeaxanthin that was undetected owing to masking by lutein.

Anthocyanin extraction and malondialdehyde assay

Anthocyanin was isolated from whole plants as described²⁹. The malondialdehyde assay for estimating lipid peroxidation was done on leaf tissue as described³⁰.

Measurement of non-photochemical quenching

Chlorophyll fluorescence in attached leaves that had been dark-adapted for 15 min was measured by a Walz PAM 101 fluorimeter. We determined NPQ by applying a light saturation pulse after illumination with actinic light over a range of intensities (500, 1,000, 2,000, 4,000 or 6,000 µmol photons m⁻² s⁻¹) for 6 min.

Measurement of leaf temperature

We measured leaf temperature on the underside of the leaf under ambient light and temperature conditions using a thermocouple (type T) and electronic thermometer (model 9009, Comark).

Received 7 January; accepted 2 May 2002; doi:10.1038/nature00861.

- Khush, G. Chair's introduction. Rice Biotechnology: improving yield, stress tolerance and grain quality. *Novartis Foundation Symp.* **236**, 11–12 (2001).
- Niyogi, K. Photoprotection revisited: genetic and molecular approaches. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 333–359 (1999).

- Cogdell, R. J. & Frank, H. A. How carotenoids function in photosynthetic bacteria. *Biochim. Biophys. Acta* **895**, 63–79 (1987).
- Yamamoto, H. Y., Nakayama, T. O. M. & Chichester, C. O. Studies on the light and dark interconversions of leaf xanthophylls. *Arch. Biochem. Biophys.* **97**, 168–173 (1962).
- Demmig-Adams, B. & Adams, W. W. Photoprotection and other responses of plants to high light stress. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **43**, 599–626 (1992).
- Cunningham, F. X. & Gantt, E. Genes and enzymes of carotenoid biosynthesis in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 557–583 (1998).
- Sandmann, G. Genetic manipulation of carotenoid biosynthesis: strategies, problems and achievements. *Trends Plant Sci.* **6**, 14–17 (2001).
- Römer, S. *et al.* Elevation of the provitamin A content of transgenic tomato plants. *Nature Biotechnol.* **18**, 666–669 (2000).
- Ye, X. *et al.* Engineering the provitamin A (β-carotene) biosynthetic pathway into carotenoid-free rice endosperm. *Science* **287**, 303–305 (2000).
- Misawa, N. *et al.* Expression of an *Erwinia* phytoene desaturase gene not only confers multiple resistance to herbicides interfering with carotenoid biosynthesis but also alters xanthophyll metabolism in transgenic plants. *Plant J.* **6**, 481–489 (1994).
- Rock, C. & Zeewaar, J. The *aba* mutant of *A. thaliana* is impaired in epoxycarotenoid biosynthesis. *Proc. Natl Acad. Sci. USA* **88**, 7496–7499 (1991).
- Rissler, H. & Pogson, B. Antisense inhibition of the β-carotene hydroxylase enzyme in *Arabidopsis* and the implications for carotenoid accumulation, photoprotection and antenna assembly. *Photosynth. Res.* **67**, 127–137 (2001).
- Sun, Z., Gantt, E. & Cunningham, F. X. Cloning and functional analysis of the β-carotene hydroxylase of *Arabidopsis thaliana*. *J. Biol. Chem.* **271**, 24349–24352 (1996).
- Ruban, A. V., Lee, P., Wentworth, M., Young, A. J. & Horton, P. Determination of the stoichiometry and strength of binding of xanthophylls in the photosystem II light harvesting complexes. *J. Biol. Chem.* **274**, 10458–10465 (1999).
- Dixon, R. A. & Paiva, N. L. Stress-induced phenylpropanoid metabolism. *Plant Cell* **7**, 1085–1097 (1995).
- Xiang, C., Werner, B., Christensen, E. & Oliver, D. The biological functions of glutathione revisited in *Arabidopsis* transgenic plants with altered glutathione levels. *Plant Physiol.* **126**, 564–574 (2001).
- Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology and biotechnology. *Plant Physiol.* **126**, 485–493 (2001).
- Horton, P., Ruban, A. V. & Walters, R. G. Regulation of light harvesting in green plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 655–684 (1996).
- Li, X.-P. *et al.* A pigment-binding protein essential for regulation of photosynthetic light harvesting. *Nature* **403**, 391–395 (2000).
- Pogson, B., Niyogi, K., Björkman, O. & DellaPenna, D. Altered xanthophyll compositions adversely affect chlorophyll accumulation and nonphotochemical quenching in *Arabidopsis* mutants. *Proc. Natl Acad. Sci. USA* **95**, 13324–13329 (1998).
- Hurry, V., Anderson, J., Chow, W. & Osmond, B. Accumulation of zeaxanthin in abscisic acid-deficient mutants of *Arabidopsis* does not affect chlorophyll fluorescence quenching or sensitivity to photoinhibition in vivo. *Plant Physiol.* **113**, 639–648 (1997).
- Havaux, M. & Niyogi, K. The violaxanthin cycle protects plants from photooxidative damage by more than one mechanism. *Proc. Natl Acad. Sci. USA* **96**, 8762–8767 (1999).
- Young, A. J. & Lowe, G. M. Antioxidant and prooxidant properties of carotenoids. *Arch. Biochem. Biophys.* **385**, 20–27 (2001).
- Verhoeven, A. S., Bugos, R. C. & Yamamoto, H. Y. Transgenic tobacco with suppressed zeaxanthin formation is susceptible to stress-induced photoinhibition. *Photosynthesis Res.* **67**, 27–39 (2001).
- Havaux, M., Tardy, F., Ravenel, J., Chanu, D. & Parot, P. Thylakoid membrane stability to heat stress studied by flash spectroscopic measurements of the electrochromic shift in intact potato leaves: influence of the xanthophyll content. *Plant Cell Environ.* **19**, 1359–1368 (1996).
- Valvekens, D., Van Montagu, M. & Van Lijsebettens, M. *Agrobacterium tumefaciens*-mediated transformation of *Arabidopsis thaliana* root explants by using kanamycin selection. *Proc. Natl Acad. Sci. USA* **85**, 5536–5540 (1988).
- Noctor, G., Rees, D., Young, A. & Horton, P. The relationship between zeaxanthin, energy-dependent quenching of chlorophyll fluorescence and trans-thylakoid pH gradient in isolated chloroplasts. *Biochem. Biophys. Acta* **1057**, 320–330 (1991).
- Porra, R. J., Thompson, W. A. & Kriedemann, P. E. Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophyll *a* and *b* extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. *Biochem. Biophys. Acta* **975**, 384–394 (1989).
- Laby, R. J., Kincaid, M. S., Kim, D. G. & Gibson, S. I. The *Arabidopsis* sugar-insensitive mutants *sis4* and *sis5* are defective in abscisic acid synthesis and response. *Plant J.* **23**, 587–596 (2000).
- Hodges, D. M., DeLong, J. M., Forney, C. F. & Prange, R. K. Improving the thiobarbituric acid-reactive substance assay for estimating lipid peroxidation in plant tissues containing anthocyanin and other interfering compounds. *Planta* **207**, 604–611 (1999).

Acknowledgements

We thank A. J. Young, E. H. Murchie, A. V. Ruban and M. Wentworth for discussions; P. Scholes for technical assistance; A. Moir for DNA sequencing and oligonucleotide synthesis; and P. Mueller-Moule and D. M. Hodges for advice on assay of lipid peroxidation. This work was supported by a grant from the UK Biotechnology and Biological Science Research Council.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.H. (e-mail: p.horton@sheffield.ac.uk).

NMR analysis of a 900K GroEL–GroES complex

Jocelyne Fiaux*, Eric B. Bertelsen†, Arthur L. Horwich† & Kurt Wüthrich*

* Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule Zürich, CH-8093 Zürich, Switzerland

† Howard Hughes Medical Institute and Department of Genetics, Yale School of Medicine, New Haven, Connecticut 06510, USA

Biomacromolecular structures with a relative molecular mass (M_r) of 50,000 to 100,000 (50K–100K) have been generally considered to be inaccessible to analysis by solution NMR spectroscopy. Here we report spectra recorded from bacterial chaperonin complexes ten times this size limit (up to M_r 900K) using the techniques of transverse relaxation-optimized spectroscopy and cross-correlated relaxation-enhanced polarization transfer^{1–5}. These techniques prevent deterioration of the NMR spectra by the rapid transverse relaxation of the magnetization to which large, slowly tumbling molecules are otherwise subject. We tested the resolving power of these techniques by examining the isotope-labelled homoheptameric co-chaperonin GroES (M_r 72K), either free in solution or in complex with the homotetradecameric chaperonin GroEL (M_r 800K) or with the single-ring GroEL variant SR1 (M_r 400K). Most amino acids of GroES show

the same resonances whether free in solution or in complex with chaperonin; however, residues 17–32 show large chemical shift changes on binding. These amino acids belong to a mobile loop region of GroES that forms contacts with GroEL^{6–10}. This establishes the utility of these techniques for solution NMR studies that should permit the exploration of structure, dynamics and interactions in large macromolecular complexes.

Molecular chaperones are required for the correct folding, transport and degradation of many proteins *in vivo*. Because of these physiological functions, chaperone proteins have been the subject of intensive studies of their mechanism of action. In the *Escherichia coli* chaperonin system, the double ring-shaped homotetradecamer GroEL (M_r 800K), interacts with the dome-shaped homoheptamer GroES (M_r 72K; Fig. 1c) in the presence of ATP to form a chamber that is the site of productive polypeptide folding^{7,11–17}. The homooligomeric nature and high symmetry of GroEL, SR1, GroES and their complexes can facilitate NMR spectral analysis because the number of resonances corresponds to the size of an individual subunit, here 58K for GroEL and 10K for GroES, even though the large functional assembly is investigated.

We first measured NMR correlation spectra with uniformly ¹⁵N-labelled and perdeuterated chaperonin proteins, GroEL and SR1, using refined implementations of the principles of transverse relaxation-optimized spectroscopy (TROSY)¹ and cross-correlated relaxation-enhanced polarization transfer (CRINEPT)². In the resulting ¹⁵N–¹H correlation spectra of either SR1 or GroEL, the pattern of cross-peaks was similar and limited in number to about

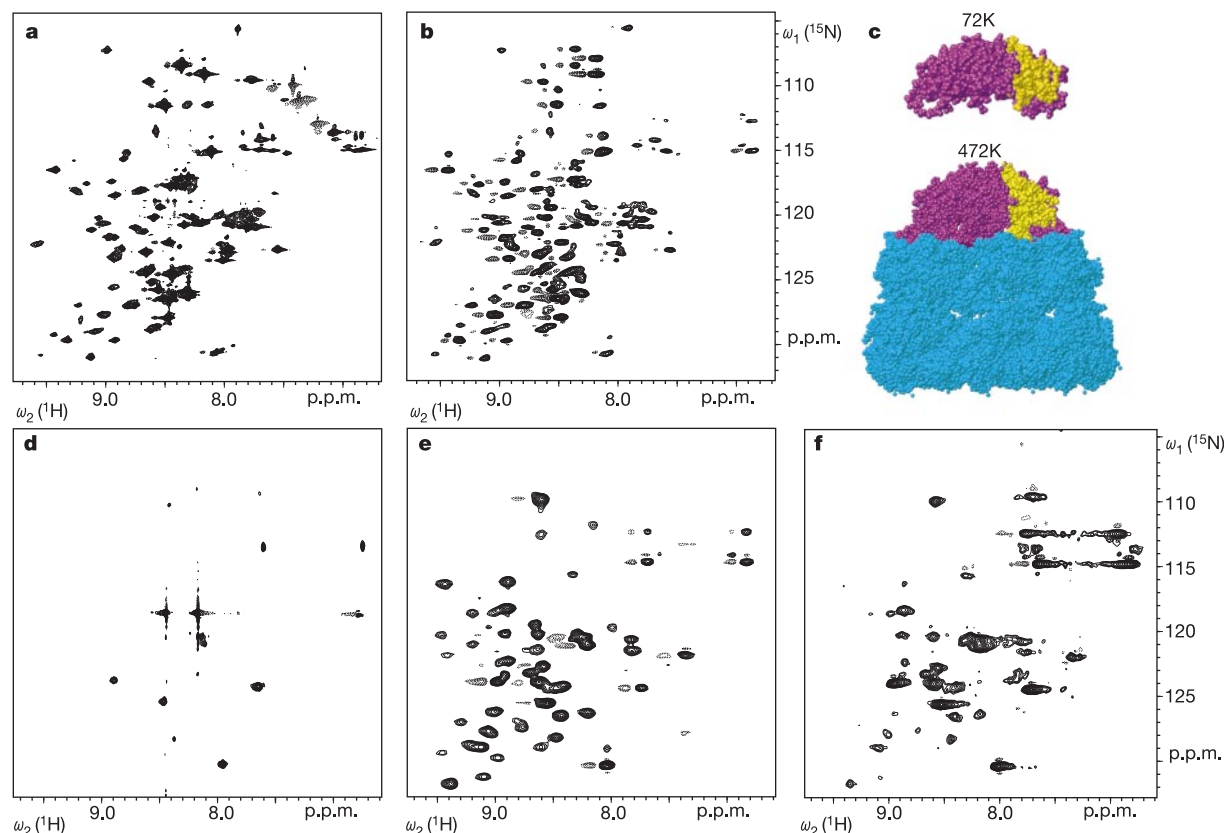


Figure 1 Two-dimensional [¹⁵N, ¹H]-correlation spectra at 25 °C of the uniformly ¹⁵N, ²H-labelled co-chaperonin GroES free in solution and in a complex with the unlabelled GroEL variant SR1. **a, b**, Free [¹⁵N, >97% ²H]GroES recorded with two-dimensional [¹⁵N, ¹H]-TROSY¹ (**a**) and two-dimensional [¹⁵N, ¹H]-CRIP-TROSY² (**b**); **c**, Top, GroES (M_r 72K) is an oligomeric protein with seven identical subunits⁶, one of which is shown in gold. Bottom, a molecular model of the complex of GroES and SR1 (M_r 472K) based on ref. 7, in which the blue-coloured SR1 is not isotope-labelled and therefore not visible in the current

NMR experiments. **d–f**, [¹⁵N, >97% ²H]GroES bound to natural isotope abundance SR1 recorded with two-dimensional [¹⁵N, ¹H]-TROSY (**d**), two-dimensional [¹⁵N, ¹H]-CRIP-TROSY (**e**) and two-dimensional [¹⁵N, ¹H]-CRINEPT-TROSY² (**f**). The concentration of free GroES was 0.3 mM in heptamer. The GroES–SR1 complex was studied in a solution containing 0.15 mM [¹⁵N, >97% ²H]GroES and 0.24 mM natural isotope abundance SR1. The experimental set-up is described in Methods.

20% of the expected resonances (data not shown). Peaks may be missing from these spectra for many reasons, including incomplete N-D/N-H exchange because the proteins were produced in perdeuterated medium, overlapping resonances, or line broadening as the result of intramolecular conformational exchange. In contrast to these results for the labelled chaperonins, virtually complete ^{15}N - ^1H correlation spectra were obtained for ^{15}N -labelled and perdeuterated GroES (Fig. 1a, b). Studies of GroES in complex with GroEL or SR1 thus provided an opportunity for unambiguous testing of the potential of TROSY/CRINEPT-based NMR techniques^{1,2} for structures with an M_r of up to 900K.

For free GroES, we detected 89 out of 94 expected backbone amide ^{15}N - ^1H cross-peaks in two-dimensional ^{15}N - ^1H -TROSY and two-dimensional ^{15}N - ^1H -CRIP-TROSY experiments (Fig. 1a, b), and we obtained sequence-specific backbone resonance assignments using TROSY triple-resonance experiments^{18–20}. Complexes of ^{15}N , ^2H GroES with chaperonin were then produced in four experiments, by its addition either to perdeuterated and ^{15}N -depleted GroEL or SR1, or to natural isotope abundance GroEL or SR1. We compared ^{15}N -depleted with natural isotope abundance chaperonin to obtain experimental proof that all the resonance lines observed in the complexes originated from GroES. Perdeuteration of the chaperonin was carried out to assess whether this would improve the observation of GroES in the complexes. No significant effect of chaperonin deuteration on the ^{15}N , ^1H GroES correlation

spectra was detected, and therefore the spectra shown were recorded using GroES complexes with unlabelled chaperonins.

Complex formation during stepwise addition of the deuterated chaperonins was evident from a decrease of the GroES TROSY cross-peak intensities (data not shown). After adding one equivalent of either SR1 or GroEL, nearly all of the TROSY resonances observed with free GroES had disappeared (Figs 1a, d and 2a, d), almost certainly because transverse relaxation in the complexes was too fast to allow observation of the NMR signals with the TROSY experimental scheme. Yet the TROSY spectra of the complexes did contain a few peaks, which we attributed to regions of GroES that remained mobile (Figs 1d and 2d). One of these resonances was identifiable as Ala 97, the carboxy-terminal residue of GroES (Fig. 2d, peak 97).

The incorporation of GroES into a much larger structure was also readily apparent in ^{15}N , ^1H -CRIP-TROSY spectra (Figs 1b, e, 2b, e, f and 3a–d). In contrast to the TROSY experiment, most of the 89 ^{15}N - ^1H cross-peaks seen in free GroES (Fig. 1b) remained resolved in the ^{15}N , ^1H -CRIP-TROSY spectra of the complexes. Moreover, for nearly all cross-peaks, only the most slowly relaxing one of the four components of the multiplet pattern (Fig. 3e) was observable (Figs 1e, 2b, e, f and 3b, d). The other three components were broadened beyond detection in the large complexes². The few cross-peaks that were also present in the ^{15}N , ^1H -TROSY spectra of the complexes (Figs 1d and 2d) showed one or two of the other multiplet components in the ^{15}N , ^1H -CRIP-TROSY spectra

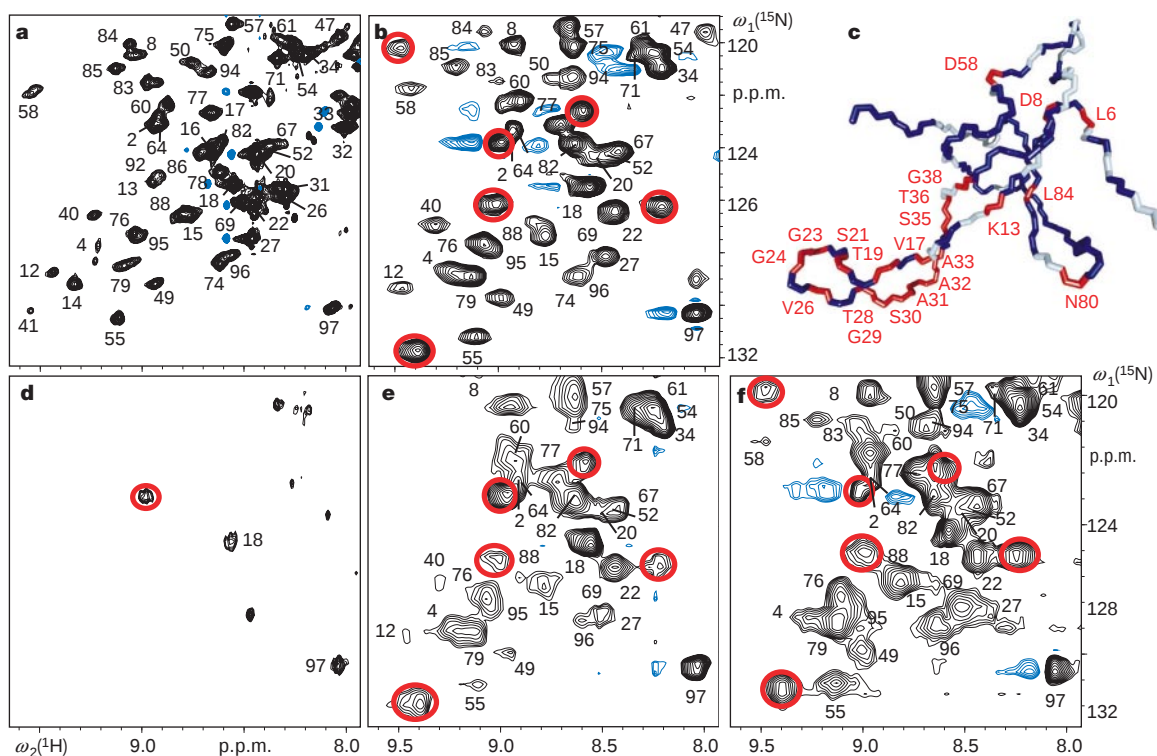


Figure 2 Chemical shift changes in $[U-^{15}\text{N}; >97\% ^2\text{H}]$ GroES on binding to natural isotope abundance SR1 or GroEL. **a**, ^{15}N , ^1H -TROSY spectrum of free GroES at 25 °C, which provides the GroES reference chemical shifts (same data set as in Fig. 1a). **b**, ^{15}N , ^1H -CRIP-TROSY spectrum of GroES bound to SR1 in the presence of ADP (same data set as in Fig. 1e). **c**, Molecular model of one subunit of GroES in the crystal structure⁷, coloured according to the chemical shift changes observed on complex formation with GroEL or SR1. The chemical shift variations were evaluated as $\Delta\delta_{av} = \{0.5 \times [\Delta\delta(^1\text{H})^2 + (0.2 \Delta\delta(^{15}\text{N}))^2]\}^{1/2}$ (ref. 21). Red indicates $\Delta\delta_{av} \geq 0.1$ p.p.m.; this includes the peaks for which no corresponding peak was assigned in bound GroES. All the residues in this class are identified by red lettering, which indicates the amino acid and the sequence position. Dark blue indicates $\Delta\delta_{av} < 0.1$ p.p.m. Light blue indicates no information because of spectral overlap or absence of the resonances in free GroES. **d**, ^{15}N , ^1H -

TROSY spectrum of GroES bound to GroEL at 25 °C. **e**, ^{15}N , ^1H -CRIP-TROSY spectrum of GroES bound to GroEL at 25 °C. **f**, Same as **e** but at 35 °C. In **a**, the blue peaks represent antiphase magnetization, which is not fully suppressed because of the short INEPT transfer delay of 3.4 ms. In **d**–**f**, the sample contained 0.13 mM $[U-^{15}\text{N}; >97\% ^2\text{H}]$ GroES and 0.15 mM natural isotope abundance GroEL. In all spectra, positive contour lines are black and negative contour lines are blue. The sequence-specific assignments for free GroES (black numbers in **a**) were obtained from TROSY triple-resonance experiments^{18–20} with a ^2H , ^{13}C , ^{15}N -labelled protein sample (unpublished data). Tentative assignments for the two complexes indicated in **b**, **e** and **f** have been obtained by transferring assignments of free GroES to nearby peaks of bound GroES. Peaks of bound GroES that appear in different chemical shift regions when compared with free GroES are circled in red and have not been assigned individually.

(Figs 1e and 2b, e, f), which is consistent with the idea that these residues are mobile in the complex. Finally, the ^{15}N - ^1H -CRINEPT-TROSY spectrum² (Fig. 1f) also contained the cross-peaks from both the structured and the more mobile regions of the chaperonin-bound GroES, but this experiment had lower sensitivity than the ^{15}N - ^1H -CRIP-TROSY scheme.

Comparison of the ^{15}N - ^1H -CRIP-TROSY spectra of GroES in the free form with that bound to either SR1 or GroEL showed that complex formation caused significant chemical shift changes for a discrete set of resonance lines, and that there were several cross-peaks in bound GroES that could not be correlated with any of the

peaks in free GroES (Fig. 2, red circles). Because nearly identical numbers of ^{15}N - ^1H correlation peaks were observed for free and bound GroES, and most cross-peaks were in very similar positions in free and chaperonin-bound GroES, we chose the following validated approach²¹ for a systematic comparison of free and bound GroES.

Each sequence-specific resonance assignment obtained for free GroES (unpublished data) was transferred to the nearest peak in the spectra of bound GroES, provided that the two peak positions were within a distance of 1.0 p.p.m. along the ^{15}N chemical shift axis and within 0.2 p.p.m. along the ^1H chemical shift axis. If no peak was observed in this tolerance range in the spectrum of the bound GroES, then no assignment was transferred from free to bound GroES. The peaks thus assigned in bound GroES are identified in Fig. 2b and d–f with numbers that indicate the sequence positions. There are several observations to be made about this result.

First, the cross-peaks assigned to residues 17–32 in free GroES could not be assigned in the complexes by the above procedure, whereas for most of the other residues the chemical shifts in free and bound GroES could be correlated. The ‘new’ cross-peaks in the complexes (Fig. 2, red circles) are thus likely to correspond to the mobile loop, that is, residues 17–32 of GroES^{6,8}.

Second, in free GroES the peaks of the mobile loop have random coil chemical shift values, which reflects a disordered conformation. By contrast, the ‘new’ peaks appearing on complex formation show a large chemical shift dispersion, which indicates that the mobile loop of GroES has become ordered on binding to GroEL. The other parts of GroES do not seem to undergo a large conformational change.

Third, most of the ^{15}N - ^1H -CRIP-TROSY cross-peaks in the spectra of the complexes consist of only the most slowly relaxing fine-structure component (Fig. 3e), which shows that the corresponding residues of GroES are immobilized relative to SR1 or GroEL, respectively, and that their brownian motions are restricted to those of the complex. But a few of the backbone ^{15}N - ^1H cross-peaks of bound GroES that are located in different positions when compared with free GroES show two or three fine-structure components in the ^{15}N - ^1H -CRIP-TROSY spectra (Figs 1e and 2b, e, f). These peaks also appear in the ^{15}N - ^1H -TROSY spectra (Figs 1d and 2d). This indicates that at least some regions of the GroES mobile loop that undergo conformational changes on binding to the chaperonins nevertheless retain significant mobility in the complex. Fourth, virtually identical results were obtained for GroES in complexes with either SR1 or GroEL (Figs 1 and 2).

These data show that only the mobile loop of GroES undergoes a large conformational change on binding to GroEL, which is compatible with previous spectroscopic and crystallographic studies of GroES interactions with GroEL^{7–10}. Additional information comes from the observation that the individual ^{15}N - ^1H cross-peaks appearing in different chemical shift positions on complex formation (Fig. 2, red circles) have variable dynamic properties, indicating a binding mode in which some of the mobile loop residues are immobilized in the complex whereas others retain a relatively high degree of mobility.

In summary, our results establish that TROSY/CRINEPT-based ^{15}N - ^1H correlation experiments^{1,2} can be used to collect informative solution NMR spectra of structures with an M_r of up to 900K. Whereas previous NMR studies on large structures have focused on regions with inherent domain mobility^{8,22–24}, our experiments can also observe structured regions. In particular, we have shown that chemical shift mapping can be used to characterize protein–protein interactions in complexes of such large size.

Further applications of this approach, with *E. coli* chaperonins and other supramolecular systems, may be facilitated greatly by the observation that deuterium labelling is needed only for the NMR-observed macromolecular component in such complexes, as shown by the fact that exactly the same cross-peaks were observed in the

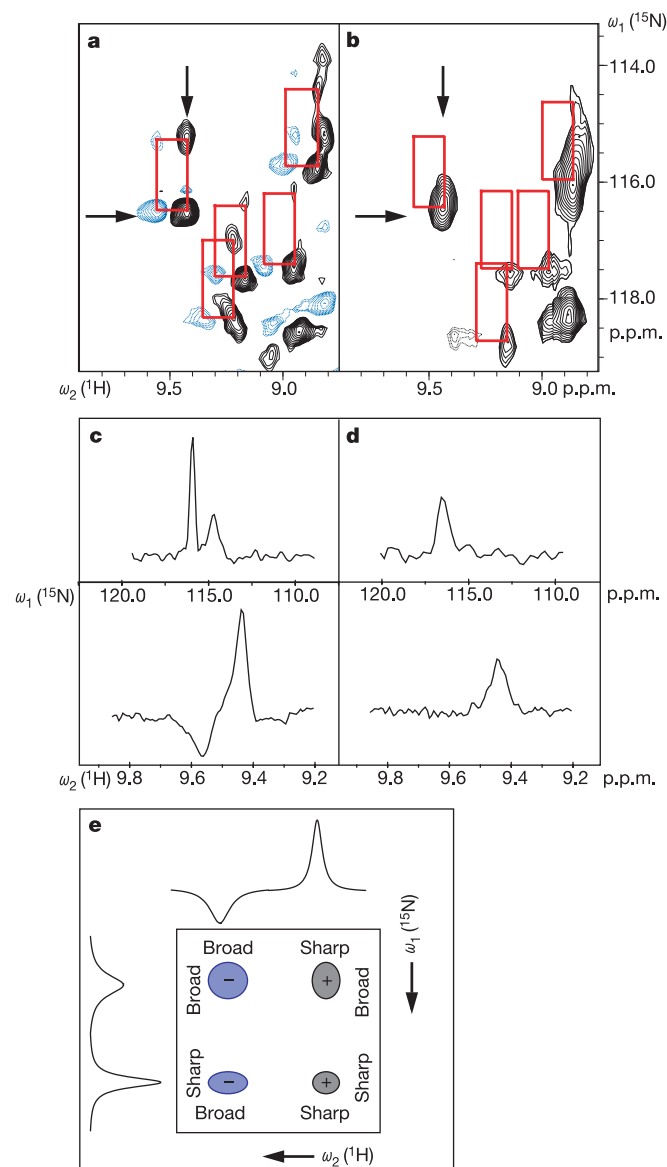


Figure 3 Multiplet patterns in ^{15}N - ^1H -CRIP-TROSY spectra of structures with an M_r of 72K and 472K. **a**, Experimental spectrum of $[\text{U-}^{15}\text{N}; >97\% \text{ } ^2\text{H}]$ GroES (same data set as in Fig. 1b). **b**, Corresponding spectral region of $[\text{U-}^{15}\text{N}; >97\% \text{ } ^2\text{H}]$ GroES bound to SR1 (same data set as in Fig. 1e). **c**, **d**, Cross-sections from **a** and **b** at the frequencies indicated by the arrows. **e**, Theoretical multiplet pattern for ^{15}N - ^1H -CRIP-TROSY cross-peaks. The four components caused by the ^1H - ^{15}N scalar coupling have different linewidths, which reflect their different transverse relaxation rates. Because the ^1H magnetization is detected in antiphase, the two components along the horizontal ^1H dimension have opposite signs, whereas the two components along the ^{15}N dimension have equal signs. For very large structures all but the ‘sharp/sharp’ fine-structure component are broadened beyond detection because of fast relaxation. In all spectra, positive contour lines are black and negative contour lines are blue.

^{15}N - ^1H correlation spectra of complexes with deuterated or protonated (unlabelled) chaperonin. The high amount of protonation in the unlabelled chaperonins also ensures that signals from these components are suppressed^{1,2}. For the *E. coli* chaperonin system, the observation that conformational changes in GroES associated with its binding to GroEL are manifested in the solution NMR spectra opens avenues for studying the functional interactions in this system, particularly for substrate polypeptide bound to GroEL. More generally, other large macromolecular assemblies should be accessible to the types of study carried out here, which would allow the interactions and dynamics of selectively labelled subunits or regions to be examined. This may considerably expand the scope of our understanding of macromolecular machines that carry out many essential cellular functions. □

Methods

Protein preparation

GroEL and GroES were expressed from a pET expression vector in strain BL21(DE3). Cells were grown on a minimal medium in 99% $^2\text{H}_2\text{O}$ supplemented with $^{15}\text{NH}_4\text{Cl}$ or $^{14}\text{NH}_4\text{Cl}$ as the sole nitrogen source, and either $^2\text{H}_3\text{[acetate]}$ for the fully deuterated ^{15}N -labelled proteins or ^{13}C glucose for the ~80% deuterated ^{13}C , ^{15}N -labelled proteins as the sole carbon source. We induced protein expression by adding 1 mM isopropylthiogalactoside in the mid-exponential growth phase and collected the cells after 8–12 h.

Cells expressing isotope-labelled GroEL were resuspended in extraction buffer I (H_2O solution of 100 mM Tris-HCl, pH 8.1, 10 mM dithiothreitol (DTT), 0.1 mM EDTA and 0.1 mg ml⁻¹ DNase I) and lysed by sonication. After adding 30% ammonium sulphate the lysate was centrifuged. We precipitated the soluble fraction with 80% ammonium sulphate and dissolved it in chromatography buffer (H_2O solution of 50 mM Tris-HCl, pH 7.2, 2 mM DTT and 0.1 mM EDTA). After overnight dialysis against the same buffer, the extract was loaded onto a DEAE Sepharose fast-flow (Pharmacia) column and eluted with a 0–500 mM NaCl gradient. We concentrated the fractions containing GroEL and applied them to a Superdex 200 column (Pharmacia) equilibrated in 25 mM potassium phosphate buffer, pH 6.15, plus 20 mM KCl, which yielded 24 mg of purified fully deuterated ^{15}N GroEL from 1 l of culture in deuterated minimal medium. Unlabelled GroEL was expressed using the same system in LB medium and purified as described²⁵ with a yield of 100 mg of purified protein from 1 l of culture.

For GroES purification, we resuspended 1 g of cells (wet weight) in extraction buffer II (10 ml of B-PER bacterial protein extraction reagent (Pierce), pH 8.0, plus 10% sucrose, 1 mg ml⁻¹ DNase I, 10 mM MgCl_2 , 5 mM isopropylfluorophosphate and a cocktail of protease inhibitors) and stirred them at 4 °C for 45 min. The lysate was cleared by centrifugation, and the pellet extraction was repeated twice at 4 °C and four times at room temperature. The extraction fractions 4–7 were pooled and applied to Source Q (Pharmacia) equilibrated in 50 mM bisTris-HCl pH 6.2 and eluted with a 0–500 mM NaCl gradient. The pooled fractions were concentrated and purified on Superdex 200 in 25 mM potassium phosphate buffer at pH 6.15 containing 20 mM KCl. The yield from 1 l of culture in deuterated minimal medium was 20 mg of purified ^{15}N , ^2H GroES, or 65 mg of purified ^{13}C , ^{15}N , ^2H GroES, respectively; the larger yield for the triply labelled protein is because glucose rather than acetate was used as the carbon source.

The single-ring variant of GroEL, SR1 (ref. 11), was expressed from a pET vector in BL21(DE3) in minimal medium in 99% $^2\text{H}_2\text{O}$ supplemented with deuterated acetate and either $^{15}\text{NH}_4\text{Cl}$ or $^{14}\text{NH}_4\text{Cl}$, and purified as described¹¹, yielding 10 mg of purified labelled SR1 from 1 l of culture. Natural isotope abundance SR1 was expressed in LB medium, yielding 100 mg from 1 l of culture.

The GroES–SR1 complex was prepared by adding 4.2 mg of [^{15}N ; >97% ^2H]GroES to 34 mg of natural isotope abundance SR1 in the presence of 10 mM ADP. The GroES–GroEL complex was prepared by adding 4.4 mg of [^{15}N ; >97% ^2H]GroES to 55 mg of natural isotope abundance GroEL in the presence of 10 mM ADP.

NMR spectroscopy

TROSY/CRINEPT-based solution NMR spectroscopy has been described^{1,2}, and these experiments have been refined for use with very large structures (R. Riek, J.F., E.B.B., A.L.H. and K.W., unpublished data). Here, three experimental schemes were used for obtaining two-dimensional ^{15}N - ^1H correlation spectra with uniformly ^{15}N -labelled and perdeuterated proteins. Scheme 1, two-dimensional [^{15}N , ^1H]-TROSY includes the widely used INEPT polarization transfer²⁶ from ^1H to ^{15}N , and vice versa, and includes transverse relaxation-optimization during the ^{15}N evolution period and the ^1H detection period. It is used for structure sizes of about 50K to 150K^{4,5}, but it is not suitable for use with larger structures because of loss of magnetization during the polarization transfers. Scheme 2, two-dimensional [^{15}N , ^1H]-CRIP-TROSY includes use of TROSY as in scheme 1. The polarization transfers are based completely on CRIP-T²⁷, which is highly efficient for large structures but has very low efficiency for spin systems with short effective correlation times, and thus tends to suppress resonance lines from such spin systems. This scheme does not select among the four fine-structure components of the ^{15}N - ^1H correlation peaks (Fig. 3), and therefore has a very short overall duration and correspondingly high sensitivity. In very large structures, the undesired three fine-structure components will be suppressed by rapid transverse relaxation (Fig. 3). Scheme 3, two-dimensional [^{15}N , ^1H]-CRINEPT-TROSY includes use of TROSY as in scheme 1. The polarization transfers are based on a combination of CRIP-T and INEPT. TROSY is active during the transfers, and

the experiment includes a selection scheme for components of the cross-peak four-line fine structure. As compared with scheme 2, scheme 3 has lower sensitivity but yields a more complete spectrum because slowly relaxing resonances are not suppressed.

All NMR samples were prepared in H_2O solutions of 25 mM potassium phosphate, pH 6.15, plus 20 mM KCl, 5% $^2\text{H}_2\text{O}$, and 0.02% NaN_3 . The NMR experiments were carried out on a Bruker DRX 750 spectrometer at a ^1H resonance frequency of 750 MHz. The sample temperature was either 25 °C or 35 °C for some experiments with the GroEL complexes. The spectra were processed with the program PROSA²⁸ and analysed with the program XEASY²⁹.

The [^{15}N , ^1H]-TROSY spectrum of free [^{15}N , ^2H]GroES (Figs 1a and 2a) was recorded for 3 h with a data size of $150 \times 1,024$ complex points, $t_{1\text{max}} = 37.5$ ms, $t_{2\text{max}} = 98$ ms, INEPT transfer time = 3.4 ms. The corresponding [^{15}N , ^1H]-CRIP-TROSY experiment (Figs 1b and 3a) was recorded for 10.5 h, with a data size of $100 \times 1,024$ complex points, $t_{1\text{max}} = 25$ ms, $t_{2\text{max}} = 98$ ms, CRIP-T transfer time = 2.8 ms. The recycle delay was 1 s. The [^{15}N , ^1H]-TROSY (Fig. 1d) and [^{15}N , ^1H]-CRIP-TROSY (Figs 1e, 2b and 3b) spectra for the GroES–SR1 complex were recorded with the same parameters as the corresponding experiments for free GroES. The [^{15}N , ^1H]-CRINEPT-TROSY experiment (Fig. 1f) was recorded for 10.5 h, with a data size of $100 \times 1,024$ complex points, $t_{1\text{max}} = 25$ ms, $t_{2\text{max}} = 98$ ms, CRINEPT transfer time = 2.8 ms. A recycle delay of 0.6 s was used for the three experiments with the GroES–SR1 complex.

For the GroES–GroEL complex, the [^{15}N , ^1H]-TROSY spectrum (Fig. 2d) was recorded for 14 h, with a data size of $120 \times 1,024$ complex points, $t_{1\text{max}} = 30$ ms, $t_{2\text{max}} = 98$ ms. The INEPT transfer time was 3.4 ms. The [^{15}N , ^1H]-CRIP-TROSY spectra (Fig. 2e, f) were recorded for 15 h each, with a data size of $64 \times 1,024$ complex points, $t_{1\text{max}} = 16$ ms, $t_{2\text{max}} = 98$ ms, recycle delay = 0.3 s. The CRIP-T transfer time was 1.4 ms. For all [^{15}N , ^1H]-CRIP-TROSY and [^{15}N , ^1H]-CRINEPT-TROSY data, a sine bell function shifted by 10° (ref. 30) was applied before Fourier transformation in the indirect dimension, and an empirically optimized exponential function was used in the direct dimension.

Received 28 February; accepted 30 April 2002; doi:10.1038/nature00860.

- Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated T_2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl Acad. Sci. USA* **94**, 12366–12371 (1997).
- Riek, R., Wider, G., Pervushin, K. & Wüthrich, K. Polarization transfer by cross-correlated relaxation in solution NMR with very large molecules. *Proc. Natl Acad. Sci. USA* **96**, 4918–4923 (1999).
- Wüthrich, K. *NMR of Proteins and Nucleic Acids* (Wiley, New York, 1986).
- Wider, G. & Wüthrich, K. NMR spectroscopy of large molecules and multi-molecular assemblies in solution. *Curr. Opin. Struct. Biol.* **9**, 594–601 (1999).
- Riek, R., Pervushin, K. & Wüthrich, K. TROSY and CRINEPT: NMR with large molecular and supramolecular structures in solution. *Trends Biochem. Sci.* **25**, 462–468 (2000).
- Hunt, J. F., Weaver, A. J., Landry, S. J., Gierasch, L. & Deisenhofer, J. The crystal structure of the GroES co-chaperonin at 2.8 Å resolution. *Nature* **379**, 37–45 (1996).
- Xu, Z., Horwich, A. L. & Sigler, P. B. The crystal structure of the asymmetric GroEL–GroES–(ADP)₇ chaperonin complex. *Nature* **388**, 741–750 (1997).
- Landry, S. J., Zeilstra-Ryalls, J., Fayet, O., Georgopoulos, C. & Gierasch, L. M. Characterization of a functionally important mobile domain of GroES. *Nature* **364**, 255–258 (1993).
- Shewmaker, F., Maskos, K., Simmerling, C. & Landry, S. J. The disordered mobile loop of GroES folds into a defined β -hairpin upon binding GroEL. *J. Biol. Chem.* **276**, 31257–31264 (2001).
- Landry, S. J., Taher, A., Georgopoulos, C. & van der Vies, S. M. Interplay of structure and disorder in cochaperonin mobile loops. *Proc. Natl Acad. Sci. USA* **93**, 11622–11627 (1996).
- Horwich, A. L., Burston, S. G., Rye, H. S., Weissman, J. S. & Fenton, W. A. Construction of single-ring and two-ring hybrid versions of bacterial chaperonin GroEL. *Methods Enzymol.* **290**, 141–146 (1998).
- Braig, K. *et al.* The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. *Nature* **371**, 578–586 (1994).
- Sigler, P. B. *et al.* Structure and function in GroEL-mediated protein folding. *Annu. Rev. Biochem.* **67**, 581–607 (1998).
- Roseman, A. M., Chen, S., White, H., Braig, K. & Saibil, H. R. The chaperonin ATPase cycle: Mechanism of allosteric switching and movements of substrate-binding domains in GroEL. *Cell* **87**, 241–251 (1996).
- Weissman, J. S., Rye, H. S., Fenton, W. A., Beechem, J. M. & Horwich, A. L. Characterization of the active intermediate of a GroEL–GroES-mediated protein folding reaction. *Cell* **84**, 481–490 (1996).
- Mayhew, M. *et al.* Protein folding in the central cavity of the GroEL–GroES chaperonin complex. *Nature* **379**, 420–426 (1996).
- Brinker, A. *et al.* Dual function of protein confinement in chaperonin-assisted protein folding. *Cell* **107**, 223–233 (2001).
- Salzmann, M., Pervushin, K., Wider, G., Senn, H. & Wüthrich, K. TROSY in triple resonance experiments: New perspectives for sequential assignment of large proteins. *Proc. Natl Acad. Sci. USA* **95**, 13585–13590 (1998).
- Salzmann, M., Wider, G., Pervushin, K., Senn, H. & Wüthrich, K. TROSY-type triple-resonance experiments for sequential NMR assignments of large proteins. *J. Am. Chem. Soc.* **121**, 844–848 (1999).
- Salzmann, M., Pervushin, K., Wider, G., Senn, H. & Wüthrich, K. NMR assignment and secondary structure determination of an octameric 110 kDa protein using TROSY in triple-resonance experiments. *J. Am. Chem. Soc.* **122**, 7543–7548 (2000).
- Pellecchia, M., Sebbel, P., Hermanns, U., Wüthrich, K. & Glockshuber, R. Pilus chaperone FimC–adhesin FimH interactions mapped by TROSY-NMR. *Nature Struct. Biol.* **6**, 336–339 (1999).
- Perham, R. N., Duckworth, H. W. & Roberts, G. C. K. Mobility of polypeptide chain in the pyruvate dehydrogenase complex revealed by proton NMR. *Nature* **292**, 474–477 (1981).
- McEvoy, M. M., de la Cruz, A. F. & Dahlquist, F. W. Large modular proteins by NMR. *Nature Struct. Biol.* **4**, 9 (1997).
- Arata, Y., Kato, K., Takahashi, H. & Shimada, I. Nuclear magnetic resonance study of antibodies—a multinuclear approach. *Methods Enzymol.* **239**, 440–464 (1994).
- Weissman, J. S. *et al.* Mechanism of GroEL action: productive release of polypeptide from a sequestered position under GroES. *Cell* **83**, 577–587 (1995).
- Morris, G. A. & Freeman, R. Enhancement of nuclear magnetic-resonance signals by polarization transfer. *J. Am. Chem. Soc.* **101**, 760–762 (1979).

27. Goldman, M. Interference effects in the relaxation of a pair of unlike spin-1/2 nuclei. *J. Magn. Reson.* **60**, 437–452 (1984).
28. Güntert, P., Dötsch, V., Wider, G. & Wüthrich, K. Processing of multi-dimensional NMR data with the new software PROSA. *J. Biomol. NMR* **2**, 619–629 (1992).
29. Bartels, C., Xia, T., Billeter, M., Güntert, P. & Wüthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **6**, 1–10 (1995).
30. DeMarco, A. & Wüthrich, K. Digital filtering with a sinusoidal window function: An alternative technique for resolution enhancement in FT NMR. *J. Magn. Reson.* **24**, 201–204 (1976).

Acknowledgements

This work was supported by the Schweizerischer Nationalfonds, by the Howard Hughes Medical Institute and by the NIH. We thank K. Furtak for help in constructing the plasmids used for expression of GroEL and SR1, and R. Riek for help with the NMR experiments at the outset of this project.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.W. (e-mail: wuthrich@mol.biol.ethz.ch) or A.L.H. (e-mail: horwich@csbmet.csb.yale.edu).

erratum

Optical pulsations from the anomalous X-ray pulsar 4U0142+61

B. Kern & C. Martin

Nature **417**, 527–529 (2002).

In the second sentence of this Letter, the word ‘neuron’ appears incorrectly. It should read ‘neutron’.



nature insight

RNA



Cover illustration

Electron micrograph shows nascent pre-rRNA attached to its DNA template from a lysed yeast cell. (Image courtesy of Y. Osheim, K. Wehner, A. Beyer and S. Baserga.)

The DNA molecule, as the primary repository of genetic information in living systems, is constrained to be stable and predictably structured. RNA differs little from DNA in chemical terms, but by contrast contrives to exhibit remarkable conformational flexibility and functional versatility, playing the Oscar Wilde to DNA's Marquess of Queensberry. The past few decades of intensive research have revealed, for example, that RNA physically conveys and interprets the genetic blueprint of every living cell; it performs essential structural roles in a number of molecular machines; its ability to form transient duplexes allows it to work as a switch; and it is likely to work as an essential catalyst in several biologically important reactions.

This Insight comprises an eclectic series of articles on different facets of RNA chemistry and biology. The starting point, appropriately enough, is a discussion of the pivotal role that RNA seems to have played in the evolution of life on Earth, which likely accounts for its wide distribution in present-day organisms. The catalytic properties of RNA enzymes — ribozymes — are then surveyed, and the role of RNA in ribosome structure and function is described as perhaps the best understood example of an RNA–protein assemblage.

Moving gradually from chemistry into biology, eukaryotic pre-mRNA splicing is discussed in the context of how it contributes to the diversity of proteins and, ultimately, the cells and tissues they make up. Exciting recent work on RNA-based gene regulation is then described, hinting that much remains to be learned about the role of RNA in eukaryotic regulatory networks. Finally, current efforts towards the goal of using RNA molecules as therapeutic agents are reviewed.

Space constraints mean that many fascinating aspects of RNA structure and function have had to be omitted — the organization and behaviour of certain RNA viruses for one — but we nonetheless hope that readers will find the articles stimulating and enjoyable.

Richard Turner Senior Editor

review articles:

- 214** The antiquity of RNA-based evolution
G. F. Joyce
- 222** The chemical repertoire of natural ribozymes
*J. A. Doudna
& T. R. Cech*
- 229** The involvement of RNA in ribosome function
P. B. Moore & T. A. Steitz
- 236** Alternative pre-mRNA splicing and proteome expansion in metazoans
T. Maniatis & B. Tasic
- 244** RNA interference
G. J. Hannon
- 252** Emerging clinical applications of RNA
*B. A. Sullenger
& E. Gilboa*

Editor, *Nature*: Philip Campbell
Insights Editor: Lesley Anson
Editorial Assistant: Simon Gibson
Production Editor: Simon Gribbin

Art Director: Majo Xeridat
Diagrams: Ann Thomson
Vicky Askew
Suzanne Coleman

Production: Jeffrey Bird
Marketing: Claire Aspinall
Sponsorship: Emma Jones

The antiquity of RNA-based evolution

Gerald F. Joyce

Departments of Chemistry and Molecular Biology and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA (e-mail: gjoyce@scripps.edu)

All life that is known to exist on Earth today and all life for which there is evidence in the geological record seems to be of the same form — one based on DNA genomes and protein enzymes. Yet there are strong reasons to conclude that DNA- and protein-based life was preceded by a simpler life form based primarily on RNA. This earlier era is referred to as the 'RNA world', during which the genetic information resided in the sequence of RNA molecules and the phenotype derived from the catalytic properties of RNA.

The RNA molecule has a pervasive role in contemporary biology, especially with regard to the most fundamental and highly conserved cellular processes. It is involved as a primer in DNA replication, a messenger that carries genetic information to the translation machinery, and a catalyst that lies at the heart of the ribosome. RNA instructs the processing of precursor messenger RNAs during splicing and editing, and mediates numerous other transactions of RNA and proteins in the cell. Catalytic RNAs (ribozymes) assist in RNA processing events and the replication of viral genomes. Individual nucleotides serve as important signalling molecules and their coenzyme derivatives participate in most of the reactions of central metabolism. It is as if a primitive civilization had existed prior to the start of recorded history, leaving its mark in the foundation of a modern civilization that followed. Although there may never be direct physical evidence of an RNA-based organism, because the RNA world is likely to have been extinct for almost four billion years, molecular archaeologists have uncovered artefacts of this ancestral era, none more pronounced than the recently reported crystal structure of the ribosome^{1–3}. This structure reveals the face of the RNA world in the active role that RNA has in protein synthesis.

In the laboratory, biochemists have come to appreciate the remarkable structural and functional versatility of RNA. Despite containing only four different chemical subunits, RNA folds into a variety of complex tertiary structures, analogous to structured proteins, and catalyses a broad range of chemical transformations (see review in this issue by Doudna and Cech, pages 222–228). RNA evolution in the laboratory, which can be viewed as a model of RNA evolution in the RNA world, has been used to obtain many new RNA enzymes. These include RNAs that catalyse nucleotide synthesis⁴ RNA polymerization⁵, aminoacylation of trans-

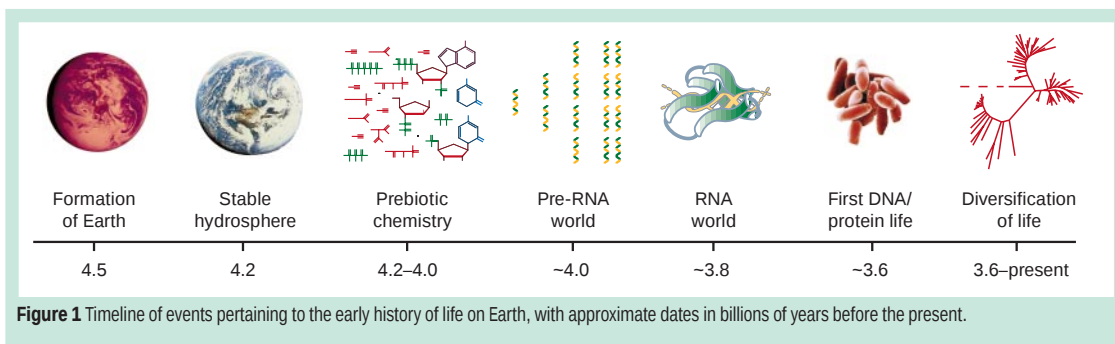
fer RNA⁶ and peptide bond formation⁷. It seems likely that RNA has the capability to support life based on RNA genomes that are copied and maintained through the catalytic function of RNA. There are substantial gaps, however, in scientific understanding concerning how the RNA world arose, the degree of metabolic complexity that it attained, and the way that it led to DNA genomes and protein enzymes.

The dawn of darwinian evolution

The worlds of prebiotic chemistry and primitive biology lie on opposite sides of the defining moment for life, when darwinian evolution first began to operate (Fig. 1). Before that time, chemical processes may have led to a substantial level of complexity. Depending on the nature of the prebiotic environment, available building blocks may have included amino acids, hydroxy acids, sugars, purines, pyrimidines and fatty acids. These could have combined to form polymers of largely random sequence and mixed stereochemistry (handedness). Some of the polymers may have had special properties, such as adherence to a particular mineral surface, unusual resistance to degradation, or the propensity to form supramolecular aggregates. Eventually every polymer, no matter how stable, would have succumbed to degradation.

A special class of polymers are those that are capable of self-replication. Although polymer self-replication is often interpreted as involving residue-by-residue copying of the polymer — a view biased by familiarity with nucleic acid replication in biology — all that is actually required is that the polymer gives rise to additional polymer molecules of the same sequence. If the rate of production of new copies exceeds the rate of degradation of existing copies, then a particular polymer sequence will persist over time.

Natural environments are subject to fluctuating conditions, ranging from diurnal and seasonal variation to



unpredictable and potentially cataclysmic events. When the environment is altered, the special properties associated with a particular polymer may no longer apply and the capacity for self-replication may be lost. Persistence in a changing environment requires a more general mechanism for self-replication that allows the polymer sequence to change somewhat over time, but retain its heritage in most of the sequence that is unchanged. The polymer must be replicated in essentially the same manner regardless of its sequence. Variation will arise owing to inevitable copying errors, and those variants too must be amenable to replication.

Once a general mechanism existed for self-replication, allowing the introduction of variation and the ability to replicate those variants, darwinian evolution began to operate. This marked the beginning of life. The special properties of a particular polymer sequence then were defined by its net rate of accumulation (rate of production minus rate of degradation), and sequences that were associated with the most favourable survival rates would have come to dominate their locale. From that point onward, the natural history of life on Earth played out as a succession of dominant polymer sequences and their associated functional properties.

A cluttered path to RNA

RNA is a polymer of variable sequence that is amenable to self-replication by a templating mechanism⁸. Different sequences have different chemical properties, but almost all sequences are able to form Watson–Crick duplex structures that facilitate the production of new copies. If the building blocks of RNA were available in the prebiotic environment, if these combined to form polynucleotides, and if some of the polynucleotides began to self-replicate, then the RNA world may have emerged as the first form of life on Earth^{9,10}. But based on current knowledge of prebiotic chemistry, this is unlikely to have been the case. Ribose, phosphate, purines and pyrimidines all may have been available, although the case for pyrimidines is less compelling^{11,12}. These may have combined to form nucleotides in very low yield^{13,14}, complicated by the presence of a much larger amount of various nucleotide analogues. The nucleotides (and their analogues) may even have joined to form polymers, with a combinatorial mixture of 2',5'-, 3',5'- and 5',5'-phosphodiester linkages, a variable number of phosphates between the sugars, D- and L- stereoisomers of the sugars, α - and β -anomers at the glycosidic bond, and assorted modifications of the sugars, phosphates and bases (Fig. 2). It is difficult to visualize a mechanism for self-replication that either would be impartial to these compositional differences or would treat them as sequence information in a broader sense and maintain them as heritable features.

The chief obstacle to understanding the origin of RNA-based life is identifying a plausible mechanism for overcoming the clutter wrought by prebiotic chemistry. Several avenues of investigation are

being pursued. Perhaps there were special conditions that led to the preferential synthesis of activated β -D-nucleotides or the preferential incorporation of these monomers into polymers. For example, the prebiotic synthesis of sugars from formaldehyde can be biased by starting from glycoaldehyde phosphate, leading to ribose 2,4-diphosphate as the predominant pentose sugar¹⁵. This reaction can occur starting from dilute aqueous solutions of reactants at near-neutral pH when carried out in the presence of certain metal-hydroxide minerals¹⁶. The polymerization of adenylate, activated as the 5'-phosphorimidazole, yields 2',5'-linked products in solution, but mostly 3',5'-linked products in the presence of a montmorillonite clay¹⁷. Thus, through a series of biased syntheses, fractionations and other enrichment processes, there may have been a special route to a warm little pond of RNA.

Another approach is to hypothesize that life did not begin with RNA; some other genetic system preceded RNA, just as it preceded DNA and protein (Fig. 1). This approach has met with substantial progress in recent years, despite the lack of guidance from known metabolic pathways in biology regarding the chemical nature of a precursor to RNA. A systematic investigation of potentially natural nucleic acid analogues containing various sugars and linkage isomers has led to the recognition of some intriguing pairing systems¹⁸. Most notable is the threose nucleic acid (TNA) analogue based on α -L-threofuranosyl units joined by 3',2'-phosphodiester linkages¹⁹ (Fig. 3a). This analogue forms stable Watson–Crick pairs with itself and with RNA. From the point of view of overcoming the clutter of prebiotic chemistry, TNA is more advantageous than RNA because of its relative chemical simplicity. Threose is one of only two aldotetroses (four-carbon sugars) and can only be joined at the 2' and 3' positions. Additionally, it is not difficult to imagine how a 'TNA world' might have made the transition to an RNA world while preserving the continuity of genetic information (see below).

There are other interesting candidates for a potential predecessor to RNA. Peptide nucleic acid (PNA) consists of a peptide-like backbone of *N*-(2-aminoethyl)glycine units with the bases attached through a methylenecarbonyl group²⁰ (Fig. 3b). Aminoethylglycine has been synthesized in spark discharge reactions from nitrogen, ammonia, methane and water²¹, although the prebiotic synthesis of an entire PNA monomer has not been achieved. PNA forms Watson–Crick-like duplex structures with itself and with RNA. Even though it is non-chiral, PNA is susceptible to cross-inhibition of the opposing enantiomers when directing the polymerization of activated D,L-ribonucleotides^{22,23}. Furthermore, PNA monomers can undergo an intramolecular *N*-acyl transfer reaction that would prevent any conventional mechanism for their polymerization²⁴. Two other proposals for what might have come before RNA are glycerol-derived nucleic acid analogues^{25–29} (Fig. 3c) and pyranosyl-RNA (containing 4',2'-linked β -D-ribofuranosyl units)^{30,31} (Fig.

Figure 2 Prebiotic clutter surrounding RNA.

Each of the four components of RNA (coloured green, red, purple and blue) would have been accompanied by several closely related analogues (listed in black type), which could have assembled in almost any combination. All possible building blocks for each of the components should be regarded as sorting independently; for example, the phosphodiester linkage may have comprised either a 3',5' linkage involving a phosphate or a 2',5' linkage involving a pyrophosphate.

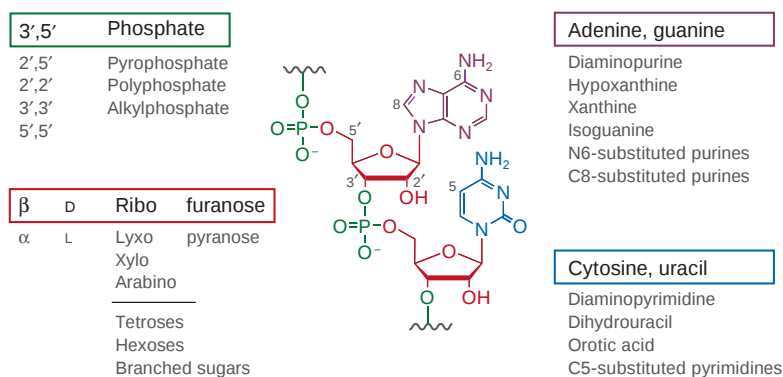
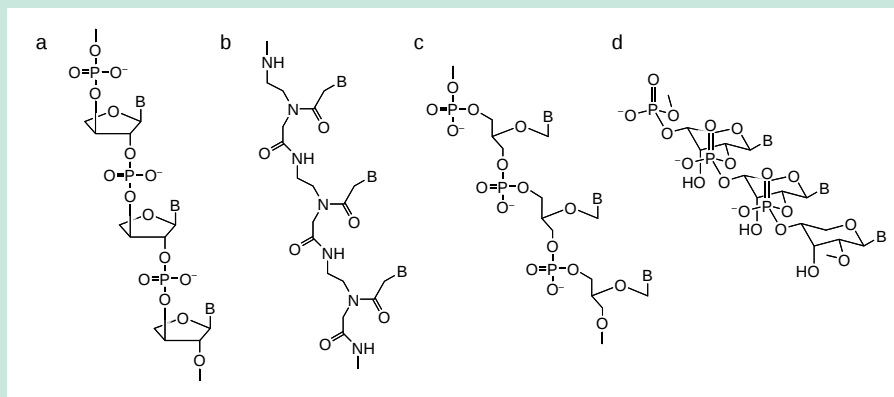


Figure 3 Candidate precursors to RNA during the early history of life on Earth. **a**, Threose nucleic acid; **b**, peptide nucleic acid; **c**, glycerol-derived nucleic-acid analogue; **d**, pyranosyl-RNA. B, nucleotide base.



3d), although neither has garnered sufficient experimental support to be considered a strong candidate.

It is also possible that RNA-based life was preceded by a replicating, evolving polymer that bore no resemblance to nucleic acids. Self-replication without darwinian evolution has been demonstrated for certain peptides³² and even small organic compounds³³. Why not cast the net broadly and consider any polymer that is capable of self-replication? A critical issue then becomes whether there is a sufficient diversity of polymer sequences that can be replicated faithfully to provide the basis for darwinian evolution. Nucleic acids have the great advantage that their potential to act as a template is sequence independent, but the templating properties of a particular nucleic acid molecule are highly sequence specific. Peptide replication based on templating within a complex of α -helices offers more restricted choices of distinct self-replicating entities, but perhaps enough to sustain a lineage of compounds in the face of a changing environment. A more radical suggestion is that the first form of life was not based on organic polymers at all, but rather on inorganic clays³⁴. Information would be represented by the distribution of charges or shapes along the surface of the clay, and replication would involve copying that information to newly formed clay layers. Suggestions of this kind challenge chemists to think more broadly about the nature of heritable chemical information and to devise experiments to test these ideas.

The transition to RNA from whatever might have preceded it would have had a very different character depending on whether the predecessor was a nucleic acid-like molecule. If the predecessor was able to cross-pair with RNA then the transition may have been a gradual one. Genetic information could have been preserved by 'transcription' of the pre-RNA to RNA, conferring selective advantage based on the function of the transcribed molecules. Once the RNA became self-replicating, it could have usurped the role of genetic material and the pre-RNA would have become expendable. If the predecessor was not a nucleic acid-like molecule, the appearance of RNA might have involved either a 'translation' process, adapting pre-RNA-based information to RNA-based information, or a 'genetic takeover'³⁵ in which none of the genetic information in pre-RNA was passed on to RNA.

Catalytic activity that resided in a pre-RNA molecule, even if that molecule was very similar to RNA, would not be expected to carry over to RNA without further evolutionary refinement. However, a pre-RNA catalyst that resembled RNA might be pre-adapted to evolve a specific function when prepared as the corresponding RNA because of preserved features of its secondary and tertiary structure. The catalytic potential of TNA, PNA and other proposed precursors to RNA has not yet been explored, but any cogent hypothesis regarding pre-RNA life must consider whether that prior genetic system could have facilitated the appearance of RNA. Once RNA appeared and became beneficial to a system undergoing darwinian evolution, further evolutionary innovation pertaining to the synthesis and

utilization of RNA would be expected to follow. In this way, pre-RNA life may have helped to overcome the problems of clutter in prebiotic synthesis by providing solutions discovered through natural selection. Eventually RNA molecules would have become responsible for ensuring the availability and replicability of RNA, ushering in the era of RNA-based darwinian evolution.

RNA-catalysed RNA replication

The general features of RNA-based life can be inferred by considering the requirements for darwinian evolution and the biochemical properties of RNA. The central process of the RNA world was the replication of RNA, presumably catalysed by RNA. The most widely studied, but by no means exclusive model for RNA replication involves template-directed polymerization of activated mononucleotides. Alternatively, replication may have involved the joining of oligonucleotides or even larger subunits³⁶, perhaps by modular assembly rather than organization along a linear template. The standard model, however, is most congruent with known biological systems and illustrates the requirements for RNA-catalysed RNA replication.

Nucleotides can be activated in several ways. The biological strategy of using nucleoside 5'-triphosphates (NTPs) is especially appealing because the linkage between the α - and β -phosphates provides a strong thermodynamic driving force (standard free energy of hydrolysis at pH 7 of about $-10 \text{ kcal mol}^{-1}$), yet is kinetically stable in typical aqueous environments ($k_{\text{hydrolysis}} \sim 10^{-10} \text{ min}^{-1}$ at pH 7 and 37°C)^{37,38}. Furthermore, polymerization of NTPs is accompanied by release of inorganic pyrophosphate, a small molecule that can readily diffuse away from the reaction centre, avoiding product inhibition. Because of its kinetic stability, however, the α,β -phosphoester reacts slowly with the 3'-hydroxyl of RNA unless the reaction is catalysed. The uncatalysed rate of joining two adjacent template-bound oligonucleotides, one bearing a 2',3'-hydroxyl and the other a 5'-triphosphate, is only $\sim 10^{-7} \text{ min}^{-1}$ at pH 7 and 37°C (ref. 39). This is comparable to the rate of hydrolysis of a single RNA phosphodiester under the same reaction conditions⁴⁰.

There is no known ribozyme in biology that catalyses the template-directed polymerization of NTPs, but such molecules have been obtained using test-tube evolution. Like the evolution of organisms in nature, evolution of RNA in the laboratory involves repeated rounds of selective amplification, linking the survival of an RNA species to its fitness⁴¹⁻⁴³. In the laboratory, fitness is defined by the experimenter, for example, based on the ability of RNA to catalyse a particular chemical reaction. Molecules that have been selected as a consequence of their function are then amplified using standard molecular biology techniques, typically reverse transcription followed by amplification using the polymerase chain reaction and then forward transcription. Random mutations may be introduced during the amplification process in order to maintain variation in the population. Through repeated rounds of selective amplification and

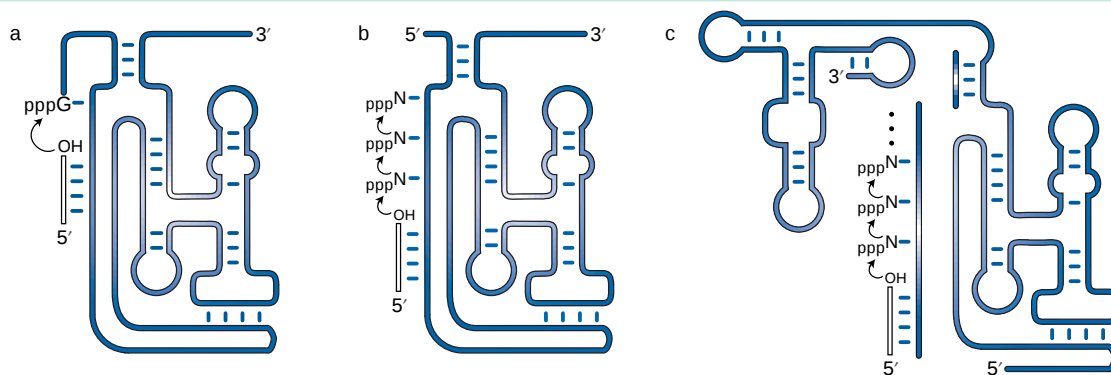


Figure 4 Successive phases in the *in vitro* evolution of an RNA polymerase ribozyme. **a**, Class I ligase ribozyme catalyzes template-directed joining of the 3' end of an RNA primer (open line) to the 5' end of the ribozyme⁴⁴. **b**, Class I ligase also catalyzes

addition of three nucleoside 5'-triphosphates (NTPs) to the 3' end of the primer, directed by an internal template⁴⁶. **c**, Class I-derived polymerase catalyzes addition of up to 14 NTPs on an external RNA template⁵.

mutation, a population of RNA molecules can be evolved to perform a defined task, provided that they have the capacity to do so.

There are several examples of *in vitro*-evolved ribozymes that catalyse the template-directed joining of an oligonucleotide 3'-hydroxyl and oligonucleotide 5'-triphosphate. The best studied of these is the class I ligase, first isolated almost ten years ago⁴⁴ (Fig. 4a). It contains an internal template region that binds a complementary RNA substrate, and directs attack of the 3'-hydroxyl of the substrate on the 5'-triphosphate of the ribozyme, forming a 3',5'-phosphodiester⁴⁵. The ribozyme contains ~120 nucleotides and operates with a catalytic rate of ~100 min⁻¹, corresponding to a rate enhancement of ~10⁹-fold compared to the uncatalysed reaction. The class I ligase also catalyses the polymerization of NTPs, adding up to three residues to the 3' end of an RNA primer in a template-directed manner⁴⁶ (Fig. 4b).

The class I ligase was used as a starting point for further evolution experiments, resulting in a ribozyme with much more robust NTP polymerization activity⁵ (Fig. 4c). The final evolved ribozyme contains ~200 nucleotides and catalyses extension of an RNA primer on an external RNA template, adding up to 14 successive nucleotides in 24 hours. It is general with respect to the template sequence, yet operates with an average fidelity of ~97% per nucleotide in copying the template sequence to that of a complementary product. This activity is not sufficient to support the RNA-catalysed replication of RNAs that are as large as the catalyst itself. However, there does not seem to be any fundamental obstacle to achieving the required level of activity, provided there exists a sufficiently powerful evolutionary search procedure.

There are likely to be many different RNA molecules that are capable of catalysing the template-directed polymerization of NTPs. The hc ligase ribozyme, also obtained by *in vitro* evolution⁴⁷, has no significant structural or sequence similarity to the class I ligase, but also catalyses the extension of an RNA primer on an external RNA template⁴⁸. It adds only one or two nucleotides, but so far has been selected for ligation rather than polymerization activity. Two other *in vitro*-evolved ribozymes, the L1 and R3 ligases, catalyse formation of a 3',5'-phosphodiester on an internal, but not external, template^{49,50}. The latter ribozyme is notable because it contains only three of the four nucleotides, completely lacking cytidine. Other ligases have been obtained that catalyse formation of a 2',5'- rather than 3',5'-phosphodiester, the ligases themselves being composed of 3',5'-linked RNA^{45,51}. A thorough exploration of the vast number of possible RNA sequences would be expected to produce numerous 3',5' ligases, many of which could be evolved into NTP polymerases. Each of these would be tolerant of substantial sequence variation by replacing base pairs in stem regions with other Watson-Crick pairs and substituting non-critical residues within loop regions by different nucleotides.

Although a very large number of RNA polymerase ribozymes might be possible, collectively they would comprise only a tiny fraction of the huge number of possible RNA sequences. For RNA molecules that contain 100 nucleotides, there are 4¹⁰⁰ (~10⁶⁰) possible sequences. A pool of one copy each of these molecules would have a mass greater than 10¹³ times that of the Earth. A pool of one copy each of all possible 40mers, with a mass of 26 kg, just might be achievable, but it is not clear if 40 nucleotides are sufficient to provide robust RNA polymerase activity. The ribozyme would be required not only to perform the chemistry of polymerization, but also to do so with sufficient fidelity to maintain the selected sequence information over successive generations. Occasional mutations are needed to maintain variability in an evolving population, but too many mutations make it impossible to retain an advantageous genotype. There is a well-established theoretical framework for assessing the effect of genome size, replication rate and replication fidelity on the ability to maintain heritable genetic information⁵². As a rule of thumb, the error rate of replication per nucleotide must be no more than about the inverse of genome length, corresponding to 99% fidelity for replication of a 100mer and 97.5% fidelity for replication of a 40mer. There may be polymerase ribozymes that meet these requirements, although such molecules have not yet been demonstrated.

The above discussion ignores other obstacles to RNA-catalysed RNA replication, such as maintaining a supply of activated mononucleotides, ensuring that the ribozyme will recognize its corresponding genomic RNA while ignoring other RNAs in the environment, overcoming stable self-structure within the template strand, separating the template and product strands, and operating in a similar manner on the product strand to generate new copies of the template. Additional genetic information might be required to overcome these obstacles, but a longer genome would necessitate an even higher fidelity of replication. Mitigating against these demands is the likelihood that RNA polymerase activity first arose in a pre-RNA world. The earliest RNA polymerases need not have been responsible for replicating entire RNA genomes, but merely for generating RNAs that enhanced the fitness of pre-RNA-based life. Further evolutionary innovation could have occurred by exploring sequences related to these functional polymerases, rather than a much broader search of all possible sequences.

Metabolic function in the RNA world

Although the central process of the RNA world was the replication of RNA genomes, some form of metabolism must have supported the process. In keeping with the second law of thermodynamics, the increase in order that occurs in a genetic system is achieved through the expenditure of high-energy starting materials that are converted

to lower-energy products. The incorporation of NTPs into an RNA polymer would qualify as a simple metabolism, although one would need to account for the source of the high-energy NTPs. Some of the starting materials may have been provided by the environment, for example, ribose and other sugars, inorganic polyphosphate, and the building blocks of purines and pyrimidines. Chemical processes in the environment may have led to more complex compounds, drawing on natural energy sources such as sunlight, electric discharges and geothermal activity. These reactions would not be considered part of metabolism, because they would not be carried out by genetically encoded catalysts. The final touches, however, leading to specific chemical organization, probably would have required the assistance of evolved catalysts. In the RNA world, those catalysts are assumed to have been ribozymes.

Ribozymes that catalyse some of the steps of nucleotide synthesis have been obtained by *in vitro* evolution (Fig. 5). One such ribozyme catalyses the formation of a nucleotide from a pyrimidine and activated ribose⁴. This is a notoriously difficult reaction in prebiotic chemistry⁵³, but was achieved starting with a pool of random-sequence RNAs that were tethered to 5-phosphoribosyl-1-pyrophosphate and allowed to react with 4-thiouracil. The evolved ribozyme performs the reaction with a catalytic rate of $\sim 0.1 \text{ min}^{-1}$ and a rate enhancement of $>10^7$ -fold compared to the uncatalysed reaction. Another *in vitro*-evolved ribozyme catalyses 5'-phosphorylation of polynucleotides, using ATP- γ -S (or ATP) as the phosphate donor⁵⁴. It operates with a catalytic rate of $\sim 0.2 \text{ min}^{-1}$ (or 0.003 min^{-1} with ATP) and a catalytic rate enhancement of $\sim 10^9$ -fold. Yet another ribozyme catalyses activation of the 5'-phosphate by attachment of a 5',5'-pyrophosphate-linked nucleotide⁵⁵. This linkage is less energetic than the α,β -phosphoanhydride of an NTP. However, ribozymes have been obtained that catalyse template-directed ligation of RNA driven by release of adenylate from a terminal adenosine-5',5'-pyrophosphate⁵⁶.

There are several important reactions in nucleotide synthesis that have not yet been carried out with a ribozyme (Fig. 5). The formation of ribose from simple aldehydes would be a significant achievement, requiring a ribozyme that catalyses a substrate-specific aldol condensation. The formation of alkylphosphates, such as glycerol phosphate or phosphorylethanolamine, would be notable, especially if the source of phosphate was a mineral or some other compound that was abundant in the environment. RNA is adept at catalysing phosphoryl transfer reactions, so it is not difficult to imagine how the phosphate, once mobilized, then would be transferred to ribose or other compounds. The synthesis of purines (for example, from cyanates) and pyrimidines (for example, from carbamoyl phosphate and aspartate) would be important as well.

The possibility of a more complex RNA-based metabolism is purely conjectural. That said, one could imagine that all of the reactions of central metabolism, now catalysed by protein enzymes, were once catalysed by ribozymes. It has been suggested that the nucleotide-derived coenzymes, which have a prominent role in most of these reactions today, are remnants of an earlier RNA-based metabolism⁵⁷. Another extreme but opposite point of view is that the only catalytic function of RNA in the RNA world was to direct the synthesis of encoded polypeptides. The synthesis of nucleotides and the template-directed polymerization of RNA may have been the responsibility of pre-RNA catalysts, with RNA merely serving as a messenger, aminoacyl adaptor and peptidyl transferase catalyst, just as it does today.

How does one assess the likelihood that a putative RNA-based function did in fact exist in the RNA world? First, it must fall within the capabilities of RNA, preferably bolstered by an experimental demonstration. Ribozymes have been obtained through *in vitro* evolution that catalyse a broad range of chemical reactions, including acyl transfer^{58,59}, N- and S-alkylation^{60,61}, carbon-carbon bond formation^{62,63}, amide bond formation⁶⁴ and Michael addition⁶⁵. But controlling a free radical within a hydrophobic pocket is likely to be

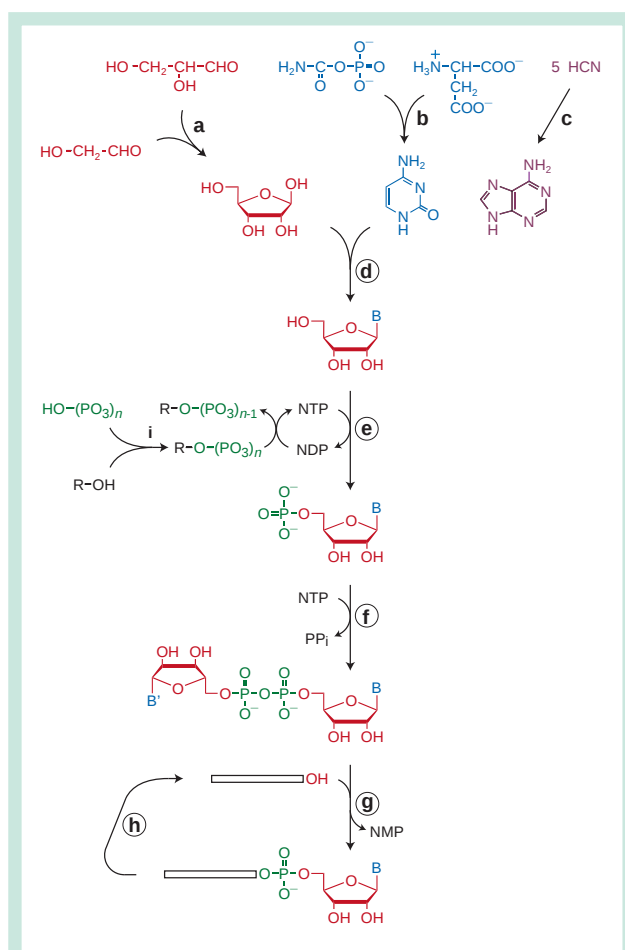


Figure 5 Hypothetical pathway for RNA-catalysed synthesis of RNA. A circled letter indicates reactions that have been demonstrated experimentally. **a**, Aldol condensation of glyceraldehyde and glycolaldehyde to form ribose. **b**, Transfer of the carbamoyl group of carbamoyl phosphate to aspartate and subsequent cyclization to form a pyrimidine. **c**, Pentamerization of HCN to form a purine. **d**, Addition of a purine or pyrimidine (B) to ribose to form a nucleoside. **e**, Phosphorylation of a nucleoside to form a nucleotide. **f**, Activation of a nucleotide by transfer of the nucleotide portion of NTP. **g**, Addition of a nucleotide to the 3' end of an RNA primer (open line). The two RNA substrates are bound at adjacent positions on a complementary template (not shown). **h**, Successive nucleotide additions resulting in further primer extension. **i**, Phosphoryl transfer from an alkyl polyphosphate to NDP, regenerating NTP. NMP is converted to NDP in a similar manner. The ultimate source of phosphate is a polyphosphate mineral.

beyond the capabilities of RNA, leading some to suggest that ribozymes never were responsible for the conversion of ribose to deoxyribose^{66,67}. Second, the RNA-catalysed reaction must have provided some selective advantage that was not contingent on future evolutionary developments. For example, a ribozyme that catalysed formation of ribosyl-1-amine from ribosyl-1-pyrophosphate and glutamine may have been selected based on the utility of the amino-sugar, but not with regard to its eventual role as an intermediate in purine biosynthesis.

If both of these requirements are met, then one should consider whether there is any evidence for the RNA-based function in contemporary biology or the geological record. It is possible, of course, that the function existed in the RNA world, but left no trace that can be detected today. Conversely, an RNA-based function that exists in contemporary biology need not have arisen in the RNA world. If, however, that function is widely distributed across all three kingdoms of life and uses RNA in a way that is not uniquely behold to the

chemical properties of RNA, then the argument that it is a remnant of the RNA world becomes more persuasive⁶⁸.

The instructed synthesis of proteins is a strong candidate for a function that existed in the RNA world. Similar rationale has been applied to suggest that tetrapyrrole biosynthesis may have arisen in the RNA world⁶⁸. The C5 pathway for the synthesis of 5-amino-levulinic acid leading to the tetrapyrroles is well represented in all three kingdoms. It uses glutamyl-transfer RNA for the synthesis of glutamate-1-semialdehyde⁶⁹, although other glutamyl esters would do just as well. In contrast, self-splicing introns are broadly distributed in contemporary biology but take special advantage of the base-pairing properties of RNA, and thus should be viewed more agnostically. Other RNA-based functions for which there is no evidence in biology, such as nucleotide synthesis and RNA polymerization, are assumed to have existed in the RNA world based on first principles, but it is important to recognize that this assumption is not supported by available historical evidence.

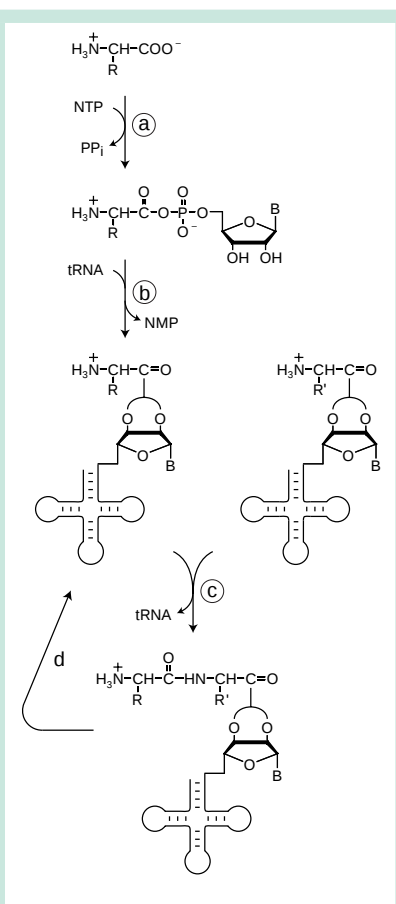
It is often said, again based on first principles rather than historical evidence, that RNA-based life must have entailed some form of cellular compartmentalization^{70,71}. This would be advantageous for keeping together an RNA replicase ribozyme and its corresponding genomic RNA, and more generally for retaining the fruits of an RNA-based metabolism for the benefit of the system that produced them. The notion of cellular compartmentalization should not be taken too literally — although all cells in contemporary biology are surrounded by a membrane composed of amphipathic lipids, there are other ways to achieve the preferential association of an ensemble of compounds. It has been proposed, for example, that there were organizing centres, analogous to modern ribosomes, where various RNAs and small molecules came together through non-covalent or transient covalent interactions⁷². Small organic molecules could have been esterified to RNA and these carrier-linked metabolites could have been held in close proximity through RNA–RNA interactions or organization along a surface. The same outcome could be achieved by passive compartmentalization on the surface of fine particulate matter, within aerosol particles in the atmosphere⁷³, or within the pores of a rock. Even if the RNA world (or pre-RNA world) synthesized compartments, those compartments might have been assembled from something other than complex phospholipids, such as nucleic acids, alternating polypeptides that form β -sheets⁷⁴, or simple terpenoids⁷⁵.

Transition to the DNA–protein world

Although RNA is well suited as a genetic molecule and can evolve to perform a broad range of catalytic tasks, it has limited chemical functionality and thus may not be equipped to meet certain challenges and opportunities that arise in the environment. An important innovation of life on Earth was the development of a separate macromolecule that would be responsible for most catalytic functions, even though that molecule contained subunits that were poorly suited for replication. The invention of protein synthesis, instructed and catalysed by RNA, was the crowning achievement of the RNA world, but also began its demise.

RNA is capable of performing all of the reactions of protein synthesis (Fig. 6). The messenger, transfer and ribosomal RNA molecules that exist in all known organisms direct the assembly of specific polypeptide sequences, instructed by corresponding RNA sequences. The activation of amino acids in the form of aminoacyl adenylates, and subsequent transfer of the amino acids to the 2'(3') terminus of tRNAs, are catalysed in modern biology by the set of 20 aminoacyl-tRNA synthetase proteins. These reactions also have been achieved with *in vitro*-evolved ribozymes. A ribozyme that contains ~110 nucleotides catalyses addition of either leucine or phenylalanine to its own 5' terminus, forming an aminoacyl-nucleotide anhydride⁷⁶. Another ribozyme catalyses aminoacylation of its own 2'(3') terminus, using various aminoacyl-adenylate substrates^{77,78}. Yet another ribozyme catalyses aminoacylation of

Figure 6 Hypothetical pathway for RNA-catalysed protein synthesis. A circled letter indicates reactions that have been demonstrated experimentally. **a**, Activation of an amino acid by formation of an aminoacyl-nucleotide anhydride. **b**, Transfer of an activated amino acid to the 2'(3') terminus of tRNA. The semicircle between the 2'- and 3'-oxygen indicates that the amino acid migrates rapidly between these two positions. **c**, Peptidyl transfer resulting in formation of a dipeptide. The two aminoacyl-tRNA substrates are bound at adjacent positions on a complementary template (not shown). **d**, Successive peptidyl transfer reactions resulting in formation of a polypeptide.



tRNAs that are either covalently attached to the ribozyme or provided as a separate substrate⁷⁹.

The final step of protein synthesis involves binding aminoacyl and peptidyl oligonucleotides at adjacent positions along an RNA template and catalysing peptide bond formation through attack of the α -amine of the amino acid on the carbonyl of the peptidyl ester. From a chemical perspective this is the easiest step, proceeding spontaneously once the reactants are brought into close proximity. It has been shown, for example, that when 2'(3')-glycyl adenosine is bound to a complementary template, peptide bond formation ensues, giving rise to diglycine, which then cyclizes to form a diketopiperazine⁸⁰. The modern ribosome achieves high template occupancy and precise orientation of the aminoacyl- and peptidyl-tRNAs, and may use additional catalytic strategies in promoting peptide bond formation⁸¹. It is not difficult to visualize how RNA alone could carry out this reaction; in fact, the crystal structure of the ribosome reveals a peptidyl transferase site that is composed entirely of RNA (ref. 82, and see review in this issue by Moore and Steitz, pages 229–235).

In vitro evolution has been used to develop ribozymes that catalyse peptide bond formation. Beginning with a pool of random-sequence RNAs with phenylalanine tethered to their 5' end, molecules were selected based on their ability to react with a methionyl-adenylate substrate to form a tethered dipeptide⁷. One of the evolved ribozymes contains ~190 nucleotides and has a catalytic rate of ~0.1 min⁻¹. It accepts several different aminoacyl adenylates, preferring methionine, and can be made to operate with multiple turnover by providing the phenylalanine substrate tethered to a short oligonucleotide rather than to the ribozyme itself⁸³. Just as the class I ligase ribozyme has been evolved to function as an RNA polymerase, this peptidyl transferase ribozyme might be evolved to form multiple peptide bonds in succession.

It is not known whether the invention of protein synthesis preceded or followed the invention of DNA genomes. The primary advantage of DNA over RNA as a genetic material is the greater chemical stability of DNA, allowing much larger genomes based on DNA. Protein synthesis may require more genetic information than can be maintained by RNA. However, the original aminoacyl adaptor molecules may have been smaller in size and fewer in number than contemporary tRNAs⁸⁴, and if the aminoacylation and peptidyl transfer activities were far less sophisticated than in the modern ribosome, an RNA genome of only a few thousand nucleotides might have been sufficient for protein synthesis. The chief argument in favour of proteins before DNA is that ribozymes seem to be incapable of catalysing the reduction of deoxyribose to ribose through the same mechanism used by all known ribonucleoside reductase proteins^{66,67}. It is conceivable, however, that RNA used a different mechanism, perhaps involving reduction of an attached purine followed by acid-catalysed elimination of the ribose 2'-hydroxyl (A. Eschenmoser, personal communication).

The template-directed polymerization of DNA is more difficult than for RNA because the 3'-hydroxyl of DNA has substantially lower acidity compared to that of RNA⁸⁵. Nonetheless, ribozymes are capable of deprotonating a DNA 3'-hydroxyl, allowing nucleophilic attack on phosphate to form a 3',5'-phosphodiester linkage⁸⁶. It is not difficult to imagine that a ribozyme could function as a DNA polymerase. Such a molecule might arise in nature or in the laboratory as an evolutionary descendant of an RNA polymerase ribozyme. RNA-based information could be reverse transcribed to DNA for safe keeping, then read back to RNA by a DNA-dependent RNA polymerase. Eventually the DNA molecules became the objects of replication, completing the transition to the DNA-protein world.

A largely open question concerns the origin of the genetic code. The aminoacylation of RNA initially must have provided some selective advantage unrelated to the eventual development of a translation machinery. It has been proposed, for example, that aminoacylation protected RNA from degradation⁸⁷, anchored RNA in an advantageous environment⁸⁷, marked genomic RNAs for replication^{87,88}, or enhanced the catalytic properties of RNA⁸⁹. Some amino acids would have been especially useful in these roles, leading to selective aminoacylation with one or perhaps a few related amino acids. Different RNAs may have been aminoacylated with different amino acids, providing the basis for a family of precursors to the modern aminoacyl-tRNAs. The RNA component of these precursor molecules may have been as simple as a stem-loop structure^{84,90} or as elaborate as a ribozyme that catalyses its own aminoacylation^{77,79}.

The next step towards the origin of the genetic code was the formation of peptide bonds between amino acids that were attached to RNA. The products of this reaction must have conferred some selective advantage, even though the peptides probably would have been too small and too heterogeneous in sequence to function as catalysts. Instead, they might have served as cofactors for ribozymes^{91,92} or been more effective than amino acids for any of the roles suggested above. RNA-catalysed peptide bond formation would have resulted in a large number of possible peptide sequences, and even this mixture may have been useful⁸⁷. However, the development of a crude mechanism for controlling the diversity of possible peptides would have been advantageous, and progressive refinement of that mechanism would have provided further selective advantage. It is reasonable to postulate that, like the modern translation apparatus, the ancestral translation system made use of messenger-like RNA molecules to gather aminoacyl-RNAs in a specific order through Watson-Crick pairing interactions. It is not clear, however, how the detailed assignments of the genetic code were made.

RNA has several features that make it suitable as the basis for a simple darwinian system: it contains only four different subunits with very similar chemical properties, its subunits polymerize readily when activated and bound to a complementary template, it is a polyanion that is readily soluble in water almost irrespective of

sequence, it forms simple secondary structures that are highly tolerant of sequence variation, and it can adopt entirely different structures following the acquisition of a few critical mutations^{93,94}. These same features make it less sophisticated compared to its DNA and protein successors. The lower reactivity but greater stability of DNA makes it a better choice for the genetic material, whereas the greater chemical diversity of the subunits of proteins, including anionic, cationic and hydrophobic groups, makes protein a better choice as the basis for catalytic function. However, those more sophisticated molecules could not have arisen without the foundation that had been laid by RNA.

Outlook

The reign of the RNA world on Earth probably began no more than about 4.2 billion years ago and ended no less than about 3.6 billion years ago⁹⁵. It may have occupied only a small portion of that interval, with the pre-RNA world having come before. Insight into the origin and operation of the RNA world is largely inferential, based on the known chemical and biochemical properties of RNA. In the best of circumstances those inferences are supported by examining the role of RNA in contemporary biology. Without that support one must be careful not to draw detailed conclusions regarding these historical events. Future studies will sharpen the picture of ancestral RNA-based life through combined efforts in prebiotic chemistry, *in vitro* evolution, biochemical analysis and molecular phylogenetics. It should be possible to formulate more precise boundary conditions regarding the environmental conditions of the early Earth and the types of chemical reactions that would have occurred under those conditions. Additional catalytic RNAs are likely to be found in biology and undoubtedly many more will be discovered through test-tube evolution. The construction of artificial RNA-based life from synthetic oligonucleotides is a distinct possibility⁷¹, and there even is a chance that a remnant of the RNA world will be found lurking in some special contemporary microenvironment. □

- Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
- Wimberly, B. T. *et al.* Structure of the 30S ribosomal subunit. *Nature* **407**, 327–338 (2000).
- Yusupov, M. *et al.* Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**, 883–896 (2001).
- Unrau, P. J. & Bartel, D. P. RNA-catalysed nucleotide synthesis. *Nature* **395**, 260–263 (1998).
- Johnston, W. K., Unrau, P. J., Lawrence, M. S., Glasner, M. E. & Bartel, D. P. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* **292**, 1319–1325 (2001).
- Lee, N., Besho, Y., Wei, K., Szostak, J. W. & Suga, H. Ribozyme-catalyzed tRNA aminoacylation. *Nature Struct. Biol.* **7**, 28–33 (2000).
- Zhang, B. & Cech, T. R. Peptide bond formation by *in vitro* selected ribozymes. *Nature* **390**, 96–100 (1997).
- von Kiedrowski, G. A self-replicating hexadeoxynucleotide. *Angew. Chem.* **25**, 932–935 (1986).
- Gilbert, W. The RNA world. *Nature* **319**, 618 (1986).
- Joyce, G. F. RNA evolution and the origins of life. *Nature* **338**, 217–224 (1989).
- Ferris, J. P., Sanchez, R. A. & Orgel, L. E. Studies in prebiotic synthesis III. Synthesis of pyrimidines from cyanoacetylene and cyanate. *J. Mol. Biol.* **33**, 693–704 (1968).
- Robertson, M. P. & Miller, S. L. An efficient prebiotic synthesis of cytosine and uracil. *Nature* **375**, 772–774 (1995).
- Lohrmann, R. & Orgel, L. E. Prebiotic synthesis: phosphorylation in aqueous solution. *Science* **161**, 64–66 (1968).
- Fuller, W. D., Sanchez, R. A. & Orgel, L. E. Studies in prebiotic synthesis. VI. Synthesis of purine nucleosides. *J. Mol. Biol.* **67**, 25–33 (1972).
- Müller, D. *et al.* Chemie von α -Aminonitrilen. Aldomerisierung von Glykolaldehydphosphat zu racemischen Hexose-2,4,6-triphosphaten und (in Gegenwart von Formaldehyd) racemischen Pentose-2,4-diphosphaten: *rac*-Allose-2,4,6-triphosphat und *rac*-Ribose-2,4-diphosphat sind die Reaktionshauptprodukte. *Helv. Chim. Acta* **73**, 1410–1468 (1990).
- Krishnamurthy, R., Pitsch, S. & Arrhenius, G. Mineral induced formation of pentose-2,4-bisphosphates. *Origins Life Evol. Biosph.* **29**, 139–152 (1999).
- Ferris, J. P. & Ertem, G. Oligomerization of ribonucleotides on montmorillonite: reaction of the 5'-phosphorimidazole of adenosine. *Science* **257**, 1387–1389 (1992).
- Eschenmoser, A. Chemical etiology of nucleic acid structure. *Science* **284**, 2118–2124 (1999).
- Schöningh, K.-U. *et al.* Chemical etiology of nucleic acid structure: the α -threofuranosyl-(3'→2') oligonucleotide system. *Science* **290**, 1347–1351 (2000).
- Nielsen, P. E., Egholm, M., Berg, R. H. & Buchardt, O. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science* **254**, 1497–1500 (1991).
- Nelson, K. E., Levy, M. & Miller, S. L. Peptide nucleic acids rather than RNA may have been the first genetic molecule. *Proc. Natl Acad. Sci. USA* **97**, 3868–3871 (2000).
- Joyce, G. F. *et al.* Chiral selection in poly(C)-directed synthesis of oligo(G). *Nature* **310**, 602–604 (1984).
- Schmidt, J. G., Nielsen, P. E. & Orgel, L. E. Enantiomeric cross-inhibition in the synthesis of oligonucleotides on a nonchiral template. *J. Am. Chem. Soc.* **119**, 1494–1495 (1997).
- Eriksson, M. *et al.* Sequence dependent N-terminal rearrangement and degradation of peptide nucleic acid (PNA) in aqueous solution. *New J. Chem.* **22**, 1055–1059 (1998).

25. Spach, G. Chiral versus chemical evolutions and the appearance of life. *Origins Life Evol. Biosph.* **14**, 433–437 (1984).
26. Joyce, G. F., Schwartz, A. W., Miller, S. L. & Orgel, L. E. The case for an ancestral genetic system involving simple analogues of the nucleotides. *Proc. Natl Acad. Sci. USA* **84**, 4398–4402 (1987).
27. Schneider, K. C. & Benner, S. A. Oligonucleotides containing flexible nucleoside analogues. *J. Am. Chem. Soc.* **112**, 453–455 (1990).
28. Merle, Y., Bonnell, E., Merle, L., Sagi, J. & Szmemo, A. Acyclic oligonucleotide analogues. *Int. J. Biol. Macromol.* **17**, 239–246 (1995).
29. Chaput, J. C. & Switzer, C. Nonenzymatic oligomerization on templates containing phosphoester-linked acyclic glycerol nucleic acid analogues. *J. Mol. Evol.* **51**, 464–470 (2000).
30. Pitsch, S., Wendeborn, S., Jaun, B. & Eschenmoser, A. Why pentose- and not hexose nucleic acids? Pyranosyl-RNA ('p-RNA'). *Helv. Chim. Acta* **76**, 2161–2183 (1993).
31. Pitsch, S. *et al.* Pyranosyl-RNA ('p-RNA'): base-pairing selectivity and potential to replicate. *Helv. Chim. Acta* **78**, 1621–1635 (1995).
32. Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K. & Ghadiri, M. R. A self-replicating peptide. *Nature* **382**, 525–528 (1996).
33. Tjivikua, T., Ballester, P. & Rebek, J. Jr A self-replicating system. *J. Am. Chem. Soc.* **112**, 1249–1250 (1990).
34. Cairns-Smith, A. G. The origin of life and the nature of the primitive gene. *J. Theor. Biol.* **10**, 53–88 (1966).
35. Cairns-Smith, A. G. & Davies, C. J. in *Encyclopaedia of Ignorance* (eds Duncan, R. & Weston-Smith, M.) 391–403 (Pergamon, Oxford, 1977).
36. James, K. D. & Ellington, A. D. The fidelity of template-directed oligonucleotide ligation and the inevitability of polymerase function. *Origins Life Evol. Biosph.* **29**, 375–390 (1999).
37. Kirby, A. J. & Younas, M. The reactivity of phosphate esters. Diester hydrolysis. *J. Chem. Soc. B* 510–513 (1970).
38. Admiral, S. J. & Herschlag, D. Catalysis of phosphoryl transfer from ATP by amine nucleophiles. *J. Am. Chem. Soc.* **121**, 5837–5845 (1999).
39. Rohatgi, R., Bartel, D. P. & Szostak, J. W. Kinetic and mechanistic analysis of nonenzymatic, template-directed oligoribonucleotide ligation. *J. Am. Chem. Soc.* **118**, 3332–3339 (1996).
40. Li, Y. & Breaker, R. R. Kinetics of RNA degradation by specific base catalysis of transesterification involving the 2'-hydroxyl group. *J. Am. Chem. Soc.* **121**, 5364–5372 (1999).
41. Robertson, D. L. & Joyce, G. F. Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344**, 467–468 (1990).
42. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
43. Ellington, A. D. & Szostak, J. W. *In vitro* selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
44. Bartel, D. P. & Szostak, J. W. Isolation of new ribozymes from a large pool of random sequences. *Science* **261**, 1411–1418 (1993).
45. Ekland, E. H., Szostak, J. W. & Bartel, D. P. Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* **269**, 364–370 (1995).
46. Ekland, E. H. & Bartel, D. P. RNA-catalysed RNA polymerization using nucleoside triphosphates. *Nature* **382**, 373–376 (1996).
47. Jaeger, L., Wright, M. C. & Joyce, G. F. A complex ligase ribozyme evolved *in vitro* from a group I ribozyme domain. *Proc. Natl Acad. Sci. USA* **96**, 14712–14717 (1999).
48. McGinness, K. E. & Joyce, G. F. RNA-catalyzed RNA ligation on an external RNA template. *Chem. Biol.* **9**, 297–307 (2002).
49. Robertson, M. P. & Ellington, A. D. *In vitro* selection of an allosteric ribozyme that transduces analytes into amplicons. *Nature Biotechnol.* **17**, 62–66 (1999).
50. Rogers, J. & Joyce, G. F. The effect of cytidine on the structure and function of an RNA ligase ribozyme. *RNA* **7**, 395–404 (2001).
51. Landweber, L. F. & Pokrovskaya, I. D. Emergence of a dual-catalytic RNA with metal-specific cleavage and ligase activities: the spandrels of RNA evolution. *Proc. Natl. Acad. Sci. USA* **96**, 173–178 (1999).
52. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523 (1971).
53. Orgel, L. E. & Lohrmann, R. Prebiotic chemistry and nucleic acid replication. *Acc. Chem. Res.* **7**, 368–377 (1974).
54. Lorsch, J. & Szostak, J. W. *In vitro* evolution of new ribozymes with polynucleotide kinase activity. *Nature* **371**, 31–36 (1994).
55. Huang, F. & Yarus, M. Versatile 5' phosphoryl coupling of small and large molecules to an RNA. *Proc. Natl Acad. Sci. USA* **94**, 8965–8969 (1997).
56. Hager, A. J. & Szostak, J. W. Isolation of novel ribozymes that ligate AMP-activated RNA substrates. *Chem. Biol.* **4**, 607–617 (1997).
57. White, H. B. III Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **7**, 101–104 (1976).
58. Lohse, P. A. & Szostak, J. W. Ribozyme-catalysed amino-acid transfer reactions. *Nature* **381**, 442–444 (1996).
59. Jenne, A. & Famulok, M. A novel ribozyme with ester transferase activity. *Chem. Biol.* **5**, 23–34 (1998).
60. Wilson, C. & Szostak, J. W. *In vitro* evolution of a self-alkylating ribozyme. *Nature* **374**, 777–782 (1995).
61. Wecker, M., Smith, D. & Gold, L. *In vitro* selection of a novel catalytic RNA: characterization of a sulfur alkylation reaction and interaction with a small peptide. *RNA* **2**, 982–994 (1996).
62. Tarasow, T. M., Tarasow, S. L. & Eaton, B. E. RNA-catalysed carbon-carbon bond formation. *Nature* **389**, 54–57 (1997).
63. Seelig, B. & Jäschke, A. A small catalytic RNA motif with Diels-Alderase activity. *Chem. Biol.* **6**, 167–176 (1999).
64. Wiegand, T. W., Janssen, R. C. & Eaton, B. E. Selection of RNA amide synthases. *Chem. Biol.* **4**, 675–683 (1997).
65. Sengle, G., Eisenführ, A., Arora, P. S., Nowick, J. S. & Famulok, M. Novel RNA catalysts for the Michael reaction. *Chem. Biol.* **8**, 459–473 (2001).
66. Freeland, S. J., Knight, R. D. & Landweber, L. F. Do proteins predate DNA? *Science* **286**, 690–692 (1999).
67. Stubbe, J., Ge, J. & Yee, C. S. The evolution of ribonucleotide reduction revisited. *Trends Biochem. Sci.* **26**, 93–99 (2001).
68. Benner, S. A., Ellington, A. D. & Tauer, A. Modern metabolism as a palimpsest of the RNA world. *Proc. Natl Acad. Sci. USA* **86**, 7054–7058 (1989).
69. Schön, A. *et al.* The RNA required in the first step of chlorophyll biosynthesis is a chloroplast glutamate tRNA. *Nature* **322**, 281–284 (1986).
70. Luisi, P. L. About various definitions of life. *Origins Life Evol. Biosph.* **28**, 613–622 (1998).
71. Szostak, J. W., Bartel, D. P. & Luisi, P. L. Synthesizing life. *Nature* **409**, 387–390 (2001).
72. Gibson, T. J. & Lamond, A. I. Metabolic complexity in the RNA world and implications for the origin of protein synthesis. *J. Mol. Evol.* **30**, 7–15 (1990).
73. Dobson, C. M., Ellison, G. B., Tuck, A. F. & Vaida, V. V. Atmospheric aerosols as prebiotic chemical reactors. *Proc. Natl Acad. Sci. USA* **97**, 11864–11868 (2000).
74. Brack, A. & Orgel, L. E. β structures of alternating polypeptides and their possible prebiotic significance. *Nature* **256**, 383–387 (1975).
75. Ourisson, G. & Nakatani, Y. The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol. *Chem. Biol.* **1**, 11–23 (1994).
76. Kumar, R. K. & Yarus, M. RNA-catalyzed amino acid activation. *Biochemistry* **40**, 6998–7004 (2001).
77. Illangasekare, M., Sanchez, G., Nickles, T. & Yarus, M. Aminoacyl-RNA synthesis catalyzed by an RNA. *Science* **267**, 643–647 (1995).
78. Illangasekare, M. & Yarus, M. Small-molecule-substrate interactions with a self-aminoacylating ribozyme. *J. Mol. Biol.* **268**, 631–639 (1997).
79. Saito, H., Kourouklis, D. & Suga, H. An *in vitro* evolved precursor tRNA with aminoacylation activity. *EMBO J.* **20**, 1797–1806 (2001).
80. Weber, A. L. & Orgel, L. E. Poly(U)-directed peptide bond formation from the 2'(3')-glycyl esters of adenosine derivatives. *J. Mol. Evol.* **16**, 1–10 (1980).
81. Barta, A. *et al.* Mechanism of ribosomal peptide bond formation. *Science* **291**, 203a (2001) (published online at www.sciencemag.org/cgi/content/full/291/5502/203a).
82. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
83. Zhang, B. & Cech, T. R. Peptidyl-transferase ribozymes: *trans* reactions, structural characterization and ribosomal RNA-like features. *Chem. Biol.* **5**, 539–553 (1998).
84. Schimmel, P. & Henderson, B. Possible role of aminoacyl-RNA complexes in noncoded peptide synthesis and origin of coded synthesis. *Proc. Natl Acad. Sci. USA* **91**, 11283–11286 (1994).
85. Izatt, R. M., Hansen, L. D., Rytting, J. H. & Christensen, J. J. Proton ionization from adenosine. *J. Am. Chem. Soc.* **87**, 2760–2761 (1965).
86. Sugimoto, N., Tomka, M., Kierzek, R., Bevilacqua, P. C. & Turner, D. H. Effects of substrate structure on the kinetics of circle opening reactions of the self-splicing intervening sequence from *Tetrahymena thermophila*: evidence for substrate and Mg^{2+} binding interactions. *Nucleic Acids Res.* **17**, 355–371 (1989).
87. Orgel, L. E. The origin of polynucleotide-directed protein synthesis. *J. Mol. Evol.* **29**, 465–474 (1989).
88. Weiner, A. M. & Maizels, N. 3' terminal tRNA-like structures tag genomic RNA molecules for replication: implications for the origin of protein synthesis. *Proc. Natl Acad. Sci. USA* **84**, 7383–7387 (1987).
89. Wong, J.-T. Origin of genetically encoded protein synthesis: a model based on selection for RNA peptidation. *Origins Life Evol. Biosph.* **21**, 165–176 (1991).
90. Schimmel, P. & Ribas de Pouplana, L. Transfer RNA: from minihelix to genetic code. *Cell* **81**, 983–986 (1995).
91. Roth, A. & Breaker, R. R. An amino acid as a cofactor for a catalytic polynucleotide. *Proc. Natl Acad. Sci. USA* **95**, 6027–6031 (1998).
92. Joyce, G. F. Nucleic acid enzymes: playing with a fuller deck. *Proc. Natl Acad. Sci. USA* **95**, 5845–5847 (1998).
93. Fontana, W. & Schuster, P. Continuity in evolution: on the nature of transitions. *Science* **280**, 1451–1455 (1998).
94. Schultes, E. A. & Bartel, D. P. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **289**, 448–452 (2000).
95. Joyce, G. F. The rise and fall of the RNA world. *New Biol.* **3**, 399–407 (1991).

Acknowledgements

I thank T. Cech, A. Eschenmoser, R. Krishnamurthy, L. Orgel, N. Paul, P. Schimmel and W. Shih for helpful comments on the manuscript. I acknowledge funding from the National Aeronautics and Space Administration and the Skaggs Institute for Chemical Biology.

The chemical repertoire of natural ribozymes

Jennifer A. Doudna* & Thomas R. Cech†

*Department of Molecular and Cell Biology, and Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA (e-mail: doudna@uclink.berkeley.edu)

†Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, Maryland 20815, USA (e-mail: thomas.cech@colorado.edu)

Although RNA is generally thought to be a passive genetic blueprint, some RNA molecules, called ribozymes, have intrinsic enzyme-like activity — they can catalyse chemical reactions in the complete absence of protein cofactors. In addition to the well-known small ribozymes that cleave phosphodiester bonds, we now know that RNA catalysts probably effect a number of key cellular reactions. This versatility has lent credence to the idea that RNA molecules may have been central to the early stages of life on Earth.

How life began on Earth is one of the great scientific mysteries. Molecular biologists have long suspected that RNA molecules were key to the process, in part because RNA has essential roles in a most fundamental process — protein synthesis — within all cells. The first example of an RNA molecule that forms a catalytic active site for a series of precise biochemical reactions was reported 20 years ago: the self-splicing pre-ribosomal RNA (rRNA) of the ciliate *Tetrahymena*. Although there was only one example, the word 'ribozyme' was coined for the general concept of an RNA molecule with enzyme-like activity¹. The following year catalytic activity was discovered in the RNA component of a ribonucleoprotein enzyme, ribonuclease (RNase) P, providing the first example of a multiple-turnover enzyme using RNA-based catalysis². These findings lent increased credibility to the hypothesis of an RNA world, where RNA served both as the genetic material and the principal cellular enzyme, probably assisted in the latter role by metal ions, amino acids and other small-molecule cofactors. The RNA world hypothesis posits that as cellular metabolism became more sophisticated, increasing demands on biocatalysts provided the impetus for the transition to protein enzymes. Descendants from this proposed RNA-dominated era inhabit today's world in the form of naturally occurring ribozymes present in organisms ranging from bacteria to humans (Table 1).

Although the known natural cellular and viral ribozymes catalyse only phosphodiester transfer chemistry, ribozymes obtained through *in vitro* selection techniques can exhibit the sort of biochemical sophistication necessary to support cellular metabolism. Starting with a pool of random RNA sequences, molecules possessing a desired activity are isolated through successive cycles of activity selection, reverse transcription of the 'winners' into DNA and amplification of those sequences by the polymerase chain reaction. This methodology has allowed identification of ribozymes that form a nucleotide from a base plus a sugar³, synthesize amide bonds^{4,5}, form Michael adducts such as those involved in the methylation of uridine monophosphate to give thymidine monophosphate⁶, and form acyl-coenzyme A, which is found in many protein enzymes⁷. It is tantalizing to think that these ribozymes are analogues of missing links in a transition from an RNA world to contemporary biology (ref. 8, and see review in this issue by Joyce, pages 214–221). Because the structures and chemical mechanisms of *in vitro*-selected

ribozymes are largely unknown at present, we focus here on the more extensively studied natural ribozymes.

RNA-based catalysis

How do RNA catalysts compare to their better-known protein enzyme counterparts? First, ribozyme rate enhancements can be substantial. For example, the rate constant for chemistry of the self-cleaving hepatitis delta virus (HDV) ribozyme is estimated at 10^2 – 10^4 s⁻¹, which is close to the maximal cleavage rate of RNase A (1.4×10^3 s⁻¹ at 25°C). The *Tetrahymena* group I self-splicing intron also has a rate constant for its chemical step that is comparable to those of protein enzymes. Second, ribozymes can use cofactors such as imidazole during catalysis^{9,10}, and they can be switched on and off by the binding of small-molecule allosteric effectors¹¹. Finally, molecular structures have revealed that ribozymes, like protein catalysts, fold into specific three-dimensional shapes that can harbour deep grooves and solvent-inaccessible active sites. These tertiary structures facilitate catalysis in part by orienting substrates adjacent to catalytic groups and metal ions.

To facilitate chemical transformations, catalysts stabilize the transition state between substrate and product. Both protein and RNA catalysts may achieve this by adding or removing protons during a reaction, orienting substrates so that they are optimally positioned to react, and using binding interactions away from the reaction site to 'force' an unfavourable contact that is relieved in the transition state. Here we discuss the structural and chemical basis for RNA catalysis as determined for several of the naturally occurring ribozymes. The lack of diverse functional groups in RNA molecules and the propensity for RNA to bind metal ions led to early hypotheses that all ribozymes might act as metalloenzymes, positioning metal ions for direct roles in catalysis. This seems to be true for group I and group II self-splicing introns and RNase P, and in the *Tetrahymena* group I intron these metal ions and their specific functions have been identified. Surprisingly, however, those ribozymes that perform site-specific strand scission — the hammerhead, hairpin, *Neurospora* Varkud satellite (VS) and HDV ribozymes — may use diverse catalytic mechanisms, and none has emerged clearly as a metalloenzyme.

Site-specific RNA self-cleavage

The hammerhead, HDV, hairpin and VS ribozymes are small RNA structures of ~40–160 nucleotides that catalyse

site-specific self-cleavage (Table 1). Found in viral, virusoid or satellite RNAs, they process the multimeric products of rolling-circle replication into genome-length strands. Although the reaction catalysed by these ribozymes is the same as that of many protein RNases (Fig. 1a), they act only at specific phosphodiester bonds by using base-pairing and other interactions to align the cleavage site within the ribozyme active site. The evolutionary maintenance of these sequences may result from the relative simplicity and efficiency of RNA-catalysed RNA strand scission.

Hammerhead ribozyme

The hammerhead, at ~40 nucleotides, is the smallest of the naturally occurring ribozymes, and mediates rolling-circle replication within circular virus-like RNAs that infect plants. Recent experiments show that the hammerhead motif (Fig. 1b) is the most efficient self-cleaving sequence that can be isolated from randomized pools of RNA, suggesting that it may have arisen multiple times during the evolution of functional RNA molecules¹². Consisting of three short helices connected at a conserved sequence junction, the hammerhead catalyses site-specific cleavage of one of its own phosphodiester bonds via nucleophilic attack of the adjacent 2'-oxygen at the scissile phosphate (Fig. 1a).

The simplicity of the hammerhead secondary structure lent itself to the design of two-piece constructs in which the strand containing the cleavage site was separated from the rest of the self-cleaving RNA. By treating one strand as the substrate and the other as the enzyme, multiple-turnover cleavage occurred with a typical rate of 1 molecule per minute at physiological salt concentrations, consistent with a substantial 10⁹-fold rate enhancement over the uncatalysed rate of nonspecific RNA hydrolysis. Initial studies revealed a requirement for a divalent metal ion for catalysis, leading to the idea that site-specific positioning of a metal ion such as magnesium might enable efficient deprotonation of the attacking 2'-hydroxyl nucleophile. Later studies at much higher ionic strength (4-M monovalent salt) showed that the hammerhead as well as the hairpin and VS ribozymes could react nearly as fast in the absence of divalent ions¹³. This discovery suggested two distinct possibilities: either these ribozymes use a different catalytic mechanism in the presence of high, non-physiological concentrations of monovalent salts, or the divalent metal ion requirement at low salt concentrations serves a structural rather than a purely chemical function.

The unveiling of the crystal structure of the hammerhead ribozyme in 1994, the first ribozyme structure to be determined, revealed a Y-shaped conformation in which nucleotides essential for catalysis were clustered at the junction of the three helical arms¹⁴ (Fig. 1c). Since then, additional crystal structures of the hammerhead

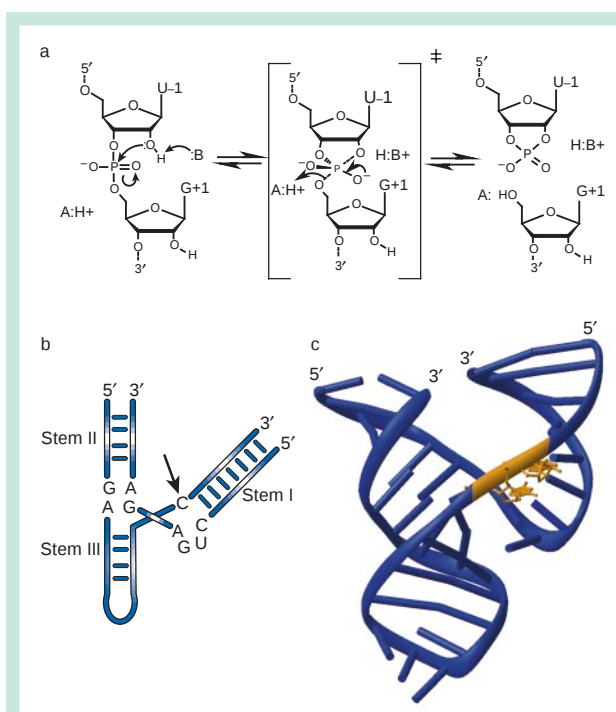


Figure 1 Mechanism of RNA-catalysed self-cleavage. **a**, General mechanism of ribonucleases and small self-cleaving ribozymes. The 2'-hydroxyl adjacent to the scissile phosphate is activated for nucleophilic attack by abstraction of its proton. Concurrently, a proton is donated to stabilize the developing negative charge on the leaving group oxygen. **b**, Secondary structure of the hammerhead ribozyme. Nucleotides important for catalytic activity are indicated; the cleavage site is indicated by an arrow. **c**, Crystal structure of the hammerhead ribozyme. Coordinates are from ref. 14. The nucleotides flanking the scissile bond are shown in gold.

ribozyme have provided 'snapshots' of the RNA at several steps along the catalytic reaction pathway: the initial pre-cleaved state, two sequential conformational changes that precede cleavage and rotate the scissile phosphate to be in line with the attacking 2'-oxygen nucleophile, and the post-cleavage product state^{15–19}. Together these structures have led to a model of ribozyme cleavage involving precise positioning of the reactive groups by the structure of the ribozyme. However, it has been difficult to ascertain from these structures what the role of bound divalent metal ions might be. Although several divalent ions were identified unambiguously in the crystal structures, they were not situated close enough to the site of catalysis to support a direct role in RNA cleavage.

Site-specific substitution of phosphate oxygens with sulphur atoms, which is readily achieved by solid-phase synthetic methods, enabled direct analysis of the effects of disrupting divalent metal-ion binding sites that were potentially involved in catalysis²⁰. These investigations led to evidence for the direct simultaneous coordination of a single metal ion by the scissile phosphate and a second phosphate oxygen located 20 Å away in the crystal structure^{21,22}. It was proposed that the crystal structures might represent the 'ground state' conformation of the hammerhead ribozyme, and that prior to catalysis the RNA conformation changed significantly but transiently to bring the critical catalytic metal ion proximal to the cleavage site. More recently, molecular modelling and kinetic analysis of the hammerhead cleavage reaction in the presence of monovalent versus divalent salts support the idea that divalent metal ions are not essential to the catalytic step, but instead stabilize the active ribozyme structure^{19,23–25}.

Whether hammerhead catalysis requires a global conformational change or merely a local rearrangement is not yet resolved. In either case, orientation of reactants within the ribozyme active site

Table 1 Naturally occurring ribozymes and ribonucleoprotein enzymes

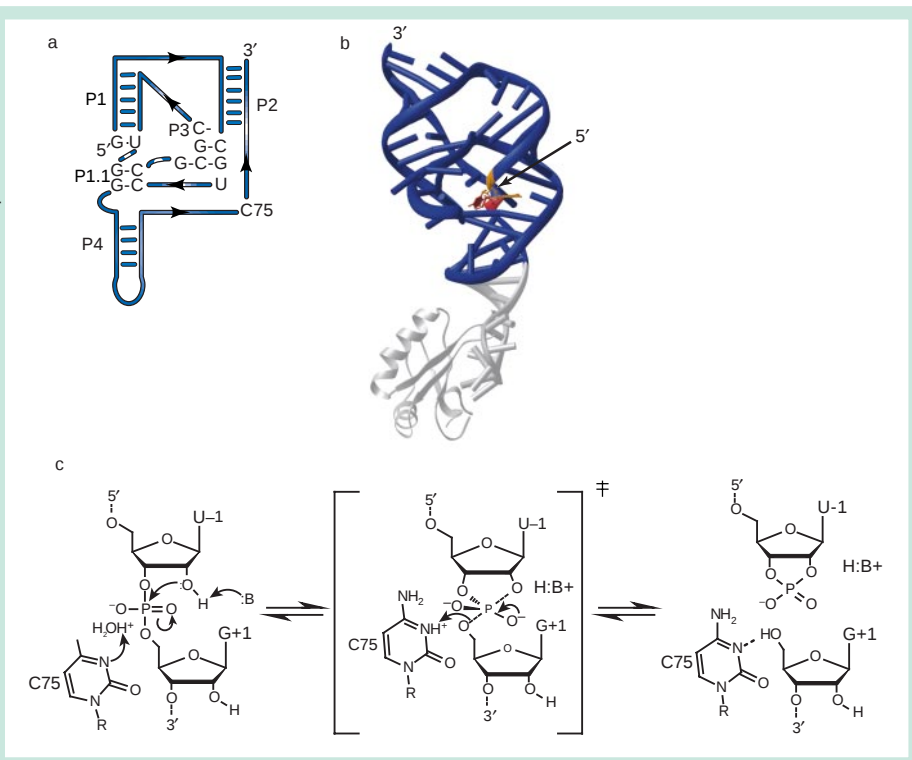
Ribozyme	Sequenced examples	Size (nt)	Activity (reaction product)
Hammerhead	11	40	Self-cleavage via transesterification (2',3' cyclic phosphate)
Hepatitis delta virus	2	90	
Hairpin	1	70	
Varkud satellite	1	160	
Group I intron	>1,500	210	Self-splicing via transesterification (3'-OH)
Group II intron	>700	500	
RNase P*	>500	300	Pre-tRNA processing via hydrolysis (3'-OH)
Spliceosome* (U2+U6 snRNAs)	70, 50	180, 100	RNA splicing via transesterification (3'-OH)
Ribosome* (23S rRNA)	>900	2,600	Peptidyl transfer (amide)

Number of sequenced examples is a snapshot as of 2002 and is influenced by DNA-sequencing strategies and database upkeep; it may provide a rough indication of relative abundance. RNAs in any group vary in size; the size provided here indicates the lower end of the length distribution for the natural examples. See www.rna.icmb.utexas.edu and www.jwbrown.mbio.ncsu.edu/RNaseP/.

*Ribonucleoprotein enzymes. RNase P: bacterial and archaeobacterial RNAs have the relevant activity in the absence of protein. Spliceosome: U2 and U6 small nuclear RNAs (snRNAs) alone show an activity related to the natural activity. Ribosome: no activity has yet been observed with protein-free, large-subunit rRNA.

Figure 2 Structure of the hepatitis delta virus

(HDV) ribozyme. **a**, Secondary structure of the genomic form of the HDV ribozyme; the cytosine residue essential to catalytic activity (C75) is indicated, and the cleavage site is marked by the 5' end of the RNA strand. **b**, Crystal structure of the product form of the HDV ribozyme. The active-site cytosine is shown in red, the 5' nucleotide of the ribozyme in gold, and the U1a RNA-binding-domain protein and its cognate RNA-binding site in grey (this has been engineered into the construct to assist crystallization). **c**, Proposed mechanism of general acid catalysis by C75, in which the protonated form of the C donates a proton to the leaving group during catalysis (compare with Fig. 1a).



probably contributes significantly to the rate of site-specific strand scission in the hammerhead. This is achieved through the unique structure of this RNA motif, conferred by the secondary structure and the presence of multiple conserved nucleotides in the active site. Whether these nucleotides provide anything else, such as general acid–base catalysis, is unknown.

Hepatitis delta virus and hairpin ribozymes

The HDV and hairpin ribozymes catalyse the same chemical reaction as that of the hammerhead, and they are likewise responsible for cleaving intermediates generated during rolling-circle replication of the HDV and a plant virus satellite RNA, respectively. Crystal structures of these ribozymes showed that in each case the RNA forms an enclosed cleft in which strand scission takes place^{26,27}. Furthermore, neither ribozyme seems to coordinate a divalent metal ion at the site of catalysis, but instead positions functionally essential nucleotides proximal to the substrate in a configuration suggesting the possibility of their direct role in catalysis.

The potential for RNA to use general acid–base chemistry during catalysis was unexpected. Functional groups within proteins that have pK_a values near neutrality, and thus can donate or accept a proton readily under physiological conditions, can act as general acids or general bases to shuttle protons during enzyme catalysis. But the lack of functional groups within RNA with pK_a values near physiological pH (6–7) means that for RNA to function in this way, one or more of its functional groups must have a pK_a significantly shifted towards neutrality. In RNA, adenine and cytosine have the potential for protonation of their ring nitrogens, N1 and N3, respectively, but the pK_a values for the free nucleosides are 3.5 and 4.2. However, A and C residues with substantially shifted pK_a values have been detected in small functional RNAs, presumably owing to the structural environment of the nucleotide^{28–32}.

In the HDV ribozyme, a cytosine base essential to catalytic function is positioned in a cleft adjacent to the site of cleavage in the RNA (Fig. 2a, b). A network of potential hydrogen bonds to this nucleotide is consistent with stabilization of a protonated form of the cytosine that might allow it to donate or accept a proton at some stage during

catalysis²⁶. This feature could be useful for mediating catalysis by pulling a proton off the attacking 2'-oxygen nucleophile, or by providing a proton to the 5'-oxygen leaving group (Fig. 1a). But does this in fact occur? Although detecting the movement of protons within an enzyme, or ribozyme, active site is difficult to achieve directly, several clever experiments have provided indirect evidence that such a mechanism is at work in the HDV ribozyme.

In one set of experiments, catalytic activity of ribozymes with point mutations at the critical cytosine was partially restored in the presence of imidazole, the side chain of histidine that readily accepts or donates a proton in many protein enzyme active sites⁹. Furthermore, the measured pK_a values of a series of restored reactions correlated with those of the imidazole analogues that promoted cleavage in the mutants³³. In a second series of experiments, kinetic isotope effects and correlation of reaction pK_a with the pK_a of different bases placed at the position of the critical C supported a direct role of this nucleotide in proton shuttling during catalysis^{34,35}. However, the pK_a of the active-site C is apparently transiently shifted during the reaction, as the shift is not detected within the product or precursor forms of the ribozyme³⁶. Together, the current data support a model in which the C acts as a general acid during the reaction to donate a proton to the 5'-bridging oxygen (Fig. 2c). A hydrated metal ion coordinated near the ribozyme active site may abstract a proton from the 2'-hydroxyl nucleophile³⁷ (Fig. 1a).

In the hairpin ribozyme the situation is less clear. Although the crystal structure of a precursor form of the RNA suggested that an active-site adenosine might adopt the role of a general acid or base²⁷ (Fig. 3a, b), nucleotide analogue interference experiments have not provided evidence for a shifted pK_a at this position³⁸. However, the lack of a requirement for divalent metal ions during hairpin ribozyme cleavage implies that a metal ion-independent mechanism is at work^{39–41}.

Self-splicing introns

Group I introns

Group I introns have been found to interrupt genes for rRNA, transfer RNA (tRNA) and messenger RNA (mRNA) in far-reaching

corners of biology, including the nuclei of protozoa, the mitochondria of fungi, the chloroplasts of algae, and bacteria and their phages. They are defined as a group by their common core secondary structure, consisting of an array of nine base-paired elements (P1–P9), and by their common mechanism of self-splicing. They accomplish splicing by a two-step transesterification mechanism initiated by an exogenous molecule of guanosine or guanosine triphosphate (GTP; Fig. 4a).

Because the excised intron still contains the active site for transesterification, it can be re-engineered to give a true catalyst that can cleave or ligate exogenous substrate molecules intermolecularly ('*in trans*'). The *Tetrahymena* version of this RNA enzyme, together with oligonucleotide substrates that can be synthesized with subtle chemical variations, has facilitated the dissection of the reaction pathway into its elemental steps. First, the RNA substrate base-pairs to an internal guide sequence within the ribozyme, forming the P1 helix. Second, specific ribose 2'-hydroxyl groups along the minor groove of the P1 helix promote its docking into the active site. Third, guanosine (or a 5'-phosphorylated analogue such as GTP) binds to the G site within P7. Fourth, the 3'-hydroxyl of the G acts as a nucleophile, cleaving the 5' splice-site phosphate with inversion of configuration. Finally, products are released; the slow release of the product bound by base pairing plus tertiary interactions is rate limiting for multiple-turnover cleavage under typical conditions⁴².

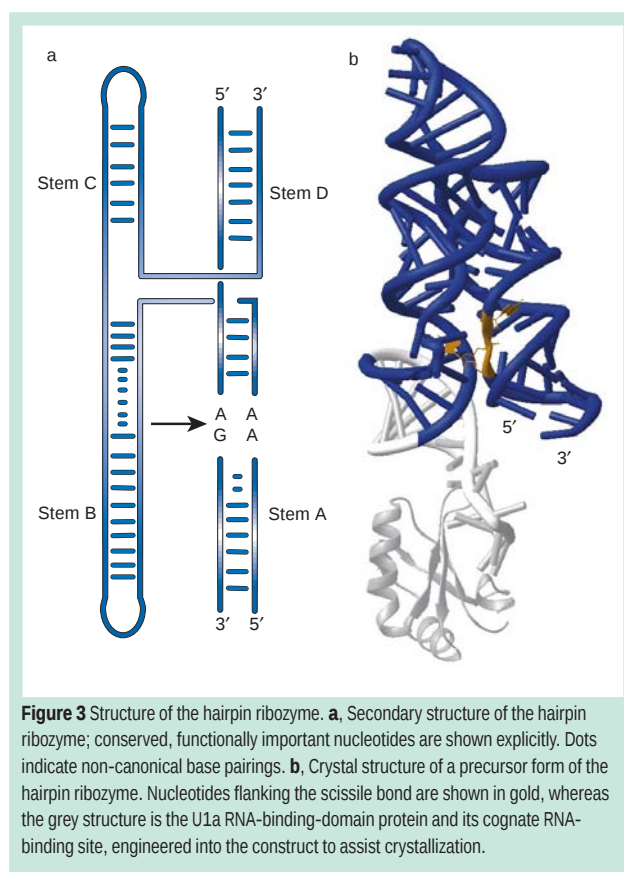
Many protein enzymes that promote phosphoryl transfer reactions use metal ions for catalysis, and group I ribozymes use the same trick⁴³. The number and location of the active-site metal ions has been investigated by substituting individual phosphate or ribose oxygen atoms with sulphur or with an amino group, and then testing for changes in metal-ion specificity (a procedure known as 'thiophilic metal-ion rescue'). The most highly supported current model is shown in Fig. 4b. One metal ion (M_B) helps deprotonate the 3'-oxygen of the G nucleophile, while another (M_A) stabilizes the developing negative charge on the leaving-group oxygen in the transition state. M_C could assist in positioning the substrates with respect to one another and, along with M_A , could help stabilize the trigonal, bipyramidal transition state⁴⁴.

The enzyme mechanism of the group I introns — involving a nucleotide-binding site, nucleophilic attack and metal-ion catalysis — is commonplace in the world of protein enzymes. But clearly the ribozyme active site must be constructed differently, given the charged and hydrophilic nature of the nucleic acid building blocks and the limited diversity of their side chains compared to amino acids. So how might a catalytic active site be built out of ribonucleotides?

The first detailed view of active-site construction came with the crystal structure of the 160-nucleotide P4–P6 domain of the *Tetrahymena* intron, an attractive target because it folds into the same structure as an excised domain as it does in the context of the whole ribozyme. This structure revealed how long-range base triples and divalent cation-mediated structures can fold an RNA molecule into a globular structure with an interior that is relatively inaccessible to solvent^{45,46}. The way in which this domain and the G-site-containing domain combine to create a concave active site was seen at modest resolution in the crystal structure of a 240-nucleotide active ribozyme⁴⁷, a structure whose general features had been predicted by modelling based on comparative phylogenetic analysis⁴⁸. Finally, the way the structure embraces the P1 substrate helix was modelled by mapping sites of chemical modification that perturb the reaction⁴⁹. Future goals are to obtain a high-resolution crystal structure of an entire group I ribozyme with bound substrates, and to locate the proposed three catalytic metals within that structure.

Group II introns

Group II introns, found in bacteria and in organellar genes of eukaryotic cells, catalyse precise self-excision and ligation of the flanking RNA sequences to form a mature transcript. In a mechanism distinct



from that of the group I introns, the group II reaction involves nucleophilic attack by the 2'-hydroxyl of a specific adenosine within the intron — the 'branch site' — to form a branched or lariat-type structure (Fig. 4a). Magnesium ions coordinated within the intron are thought to have a direct role in catalysis^{50–52}, and several studies have revealed aspects of the intron tertiary structure that are essential to catalytic function^{53–56}. Models of group II intron architecture have been proposed based on chemical probing, phylogenetic covariation and mutagenesis results^{57,58}. Interestingly, some group II introns encode proteins that assist RNA splicing and can also enable efficient integration of the intron RNA into double-stranded DNA by reverse splicing and reverse transcription⁵⁹. This reverse splicing activity promotes intron mobility by enabling insertion into targeted genes.

Ribonuclease P

RNase P, found in all cells, catalyses site-specific hydrolysis of precursor RNA substrates including tRNA, 5S rRNA and the signal recognition particle RNA^{60,61}. These substrates probably share structural features that enable efficient recognition by the RNase P substrate-binding site, positioning the reactive phosphate in each case for nucleophilic attack by a coordinated water molecule. The ribozyme is thought to be a metalloenzyme, a hypothesis supported by data from phosphorothioate substitution at the scissile phosphate of a substrate pre-tRNA^{62,63}.

RNase P is in fact an RNA–protein complex whose activity, at least in bacteria, resides within the RNA component. Furthermore, it is a true catalyst in the sense that each ribozyme complex catalyses the cleavage of multiple substrate RNAs. As with group I and II introns, much of the information about the secondary and tertiary structure of RNase P RNA has come from extensive phylogenetic covariation analysis of related sequences. Typically 300–400 nucleotides in length, it comprises two domains containing the substrate-recognition site and the ribozyme active site, respectively. Structural models

of a bacterial RNase P RNA have been proposed^{64,65}, but a crystal structure is not yet available. In human cells, the RNase P complex is larger and includes multiple protein components in addition to the RNA^{66–68}. The human RNase P RNA is not catalytically active in

the absence of protein, which has made it challenging to determine whether its active site is composed of RNA, protein or some combination of the two.

Ribozyme activity, folding and dynamics

Protein enzymes that catalyse nucleophilic attack at a phosphate within RNA or a ribonucleotide apparently utilize different chemical mechanisms depending on the enzyme. For example, mammalian adenylyl cyclases function by a two-metal-ion mechanism⁶⁹, RNase A uses two histidines for general acid–bases catalysis⁷⁰, and the anthrax adenylyl cyclase exotoxin uses one histidine and a coordinated metal ion to activate the attacking nucleophile and stabilize the leaving group, respectively⁷¹. The fact that ribozymes also catalyse phosphodiester bond cleavage by a variety of mechanisms shows that RNA has a breadth of catalytic potential similar to protein enzymes.

Furthermore, like protein enzymes, ribozymes must fold into specific three-dimensional structures to function catalytically. How do these RNAs reach their active conformations? This problem has been the subject of intense study using *in vitro* systems, and several themes are emerging. RNA structures generally fold via a cooperative, hierarchical pathway in which tertiary interactions follow the formation of a stable secondary structure. Rates of tertiary structure formation vary from tens of milliseconds to several minutes, and for the large ribozymes, can be dominated by folding intermediates that transiently trap the RNA in non-native or partially folded conformations^{72–75}. Single-molecule experiments, in which individual ribozyme molecules are analysed by fluorescence microscopy or mechanical tethering, reveal that multiple folding pathways can exist for a particular RNA sequence^{76–78}. It is not yet clear how these observations relate to RNA folding processes *in vivo*, because proteins may assist ribozyme assembly either through direct RNA binding^{79,80} or through covalent modification of specific nucleotides. In addition, RNA tertiary structures and even secondary structures can undergo rapid conformational changes^{81–85}. Such dynamics may be essential for progress through a catalytic cycle. On the other hand, too much structural flexibility may hamper catalysis, providing the incentive for evolution of ribonucleoprotein enzymes rather than pure ribozymes.

A starring role for ribozymes?

If the RNA world had a lengthy head-start over the protein catalyst world, why are RNA catalysts relatively minor players in modern cells? In fact, they may be much more central to cell biology than was previously believed. The ribosome, which is responsible for information-directed protein synthesis in all of life, is composed of three (or in some cases four) RNA molecules along with several dozen proteins. No protein subunit has ever been identified as a peptidyl transferase enzyme, and for more than 20 years, evidence for a primary role of RNA in this activity has accumulated⁸⁶. The most direct evidence came with the deduction of the crystal structure of the large subunit⁸⁷, in which the peptidyl transferase centre was precisely located by binding a small-molecule inhibitor that is an analogue of the anionic tetrahedral intermediate in amide bond formation⁸⁸. Remarkably, only RNA and no protein lies in the vicinity of the reaction centre, so the catalysis must be ribozymic. The authors suggested one possible mechanism involving a conserved adenine acting as a general base to abstract an amino proton from the amino acid⁸⁹, but subsequent mutagenesis of the key A has not provided strong support^{90,91}. Identifying the rRNA's catalytic strategy is an important direction for future research.

Given the large size of the rRNA (Table 1), scientists are using *in vitro* evolution to find smaller peptidyl transferase ribozymes that might model the biological reaction⁹². Indeed, RNA catalysts have also been identified that accomplish two other steps of the protein synthesis pathway: formation of activated amino acid adenylates⁹³ and transfer of an amino acid to the 3'-oxygen of a tRNA-like acceptor^{94,95}.

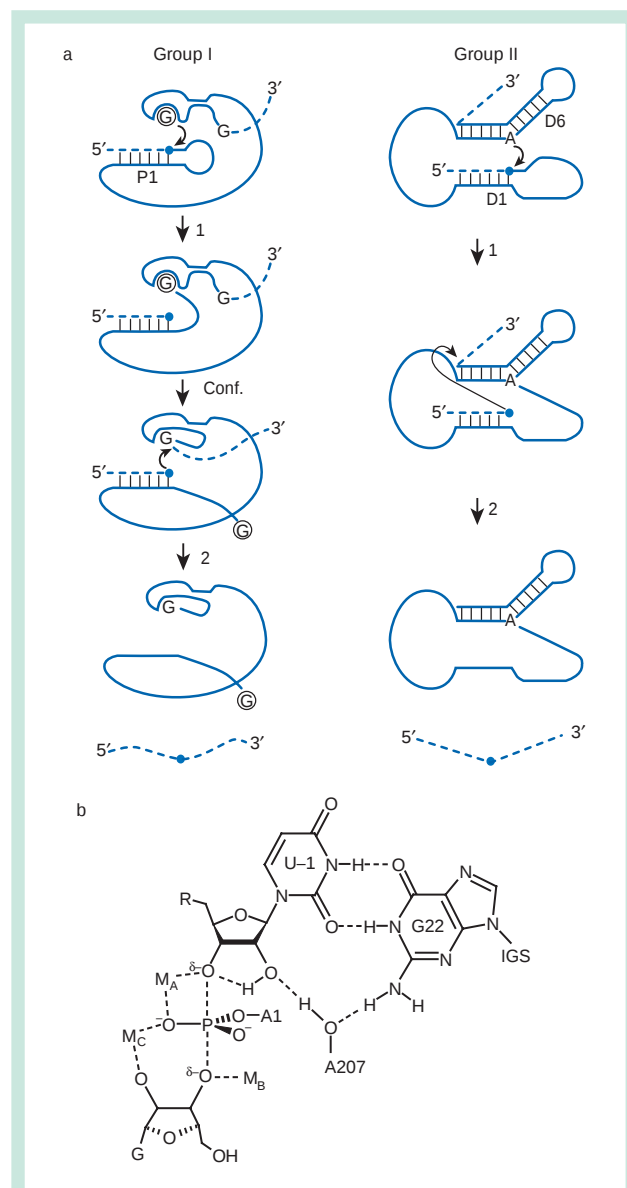


Figure 4 Self-splicing intron mechanisms. **a**, Pathways for group I and II intron self-splicing, with exons shown as dashed lines and introns as solid lines. For group I introns, step 1 shows how an intron-bound guanosine or GTP (circled) cleaves the 5' splice site while becoming covalently attached to the 5' end of the intron. 'Conf.' indicates a conformational change whereby the G at the 3' end of the intron replaces the original G in the G-binding site. In step 2, the cleaved 5' exon, still held to the intron by base pairing (P1), then cleaves the 3' splice site; as a result, the exons are ligated and the intron excised. For group II introns, step 1 shows how an adenine 2'-hydroxyl within domain 6 (D6) attacks the 5' splice site, which is identified by base-pairing interactions involving domain 1 (D1); this results in a branched 'lariat' RNA intermediate. In step 2, the cleaved 5' exon then attacks the 3' splice site, ligating the exons and excising the lariat intron. **b**, Three-metal-ion mechanism for RNA cleavage catalysed by the *Tetrahymena* group I intron (adapted from ref. 44). The step shown is the same as step 1 in **a**; the cleavage site phosphate (between U-1 and A1) is recognized in part by interactions with G22 in the internal guide sequence (IGS).

Another ribonucleoprotein catalyst found in eukaryotic cells is the spliceosome, which assembles with nuclear pre-mRNAs and splices out the major class of introns (which are not self-splicing). For many years, a popular hypothesis has been that the spliceosome is also a molecular fossil from the RNA world—with several enzymatic RNAs acting intermolecularly via a mechanism analogous to that used intramolecularly by self-splicing group II introns^{96,97}. As with the ribosome, most of the evidence for this hypothesis has been circumstantial until very recently.

Now Valadkhan and Manley⁹⁸ have shown that two spliceosomal RNAs, the U2 and U6 small nuclear RNAs, can bind an RNA substrate containing the sequence of the intron branch site and promote a splicing-related reaction in the absence of any of the numerous spliceosomal proteins. The reaction product is not the natural 'branch'—consisting of a nucleotide forming both 2'-5' and 3'-5' phosphodiester bonds—but instead a new product consistent with a phosphotriester. (This is surprising from a chemical perspective, because it would require hydroxyl as the leaving group from a pentavalent phosphorus intermediate or transition state.) This new RNA-catalysed reaction will undoubtedly stimulate fresh investigations of the mechanism by which spliceosomal RNAs catalyse mRNA splicing, and adds weight to the proposition that remnants of the RNA world are still among us.

Future directions

As some of the first RNAs to be studied in structural and mechanistic detail, ribozymes have provided many important insights into RNA function at a fundamental chemical level. Although much progress has been made, many interesting questions remain to be addressed. Determining detailed reaction mechanisms for RNA catalysts, including large RNA-protein complexes such as the ribosome and the spliceosome, will be a priority, as well as exploring the chemical mechanisms of ribozymes identified by *in vitro* selection. In addition to revealing new aspects of RNA biology, these investigations may shed light on aspects of the proposed RNA world and the role of RNA in early evolution. □

- Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* **31**, 147–157 (1982).
- Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–857 (1983).
- Unrau, P. J. & Bartel, D. P. RNA-catalysed nucleotide synthesis. *Nature* **395**, 260–263 (1998).
- Lohse, P. A. & Szostak, J. W. Ribozyme-catalysed amino-acid transfer reactions. *Nature* **381**, 442–444 (1996).
- Wiegand, T. W., Janssen, R. C. & Eaton, B. E. Selection of RNA amide synthases. *Chem. Biol.* **4**, 675–683 (1997).
- Sengle, G., Eisenfuhr, A., Arora, P. S., Nowick, J. S. & Famulok, M. Novel RNA catalysts for the Michael reaction. *Chem. Biol.* **8**, 459–473 (2001).
- Jadhav, V. R. & Yarus, M. Acyl-CoAs from coenzyme ribozymes. *Biochemistry* **41**, 723–729 (2002).
- Wilson, D. S. & Szostak, J. W. *In vitro* selection of functional nucleic acids. *Annu. Rev. Biochem.* **68**, 611–647 (1999).
- Perrotta, A. T., Shih, I. & Been, M. D. Imidazole rescue of a cytosine mutation in a self-cleaving ribozyme. *Science* **286**, 123–126 (1999).
- Santor, S. W., Joyce, G. F., Sakthivel, K., Gramatikova, S. & Barbas, C. F. III RNA cleavage by a DNA enzyme with extended chemical functionality. *J. Am. Chem. Soc.* **122**, 2433–2439 (2000).
- Tang, J. & Breaker, R. R. Rational design of allosteric ribozymes. *Chem. Biol.* **4**, 453–459 (1997).
- Salehi-Ashtiani, K. & Szostak, J. W. *In vitro* evolution suggests multiple origins for the hammerhead ribozyme. *Nature* **414**, 82–84 (2001).
- Murray, J. B., Seyhan, A. A., Walter, N. G., Burke, J. M. & Scott, W. G. The hammerhead, hairpin and VS ribozymes are catalytically proficient in monovalent cations alone. *Chem. Biol.* **5**, 587–595 (1998).
- Pley, H. W., Flaherty, K. M. & McKay, D. B. Three-dimensional structure of a hammerhead ribozyme. *Nature* **372**, 68–74 (1994).
- Scott, W. G., Finch, J. T. & Klug, A. The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell* **81**, 991–1002 (1995).
- Scott, W. G., Murray, J. B., Arnold, J. R., Stoddard, B. L. & Klug, A. Capturing the structure of a catalytic RNA intermediate: the hammerhead ribozyme. *Science* **274**, 2065–2069 (1996).
- Murray, J. B. *et al.* The structural basis of hammerhead ribozyme self-cleavage. *Cell* **92**, 665–673 (1998).
- Murray, J. B., Szoke, H., Szoke, A. & Scott, W. G. Capture and visualization of a catalytic RNA enzyme-product complex using crystal lattice trapping and X-ray holographic reconstruction. *Mol. Cell* **5**, 279–287 (2000).
- Murray, J. B., Dunham, C. M. & Scott, W. G. A pH-dependent conformational change, rather than the chemical step, appears to be rate-limiting in the hammerhead ribozyme cleavage reaction. *J. Mol. Biol.* **315**, 121–130 (2002).
- Scott, E. C. & Uhlenbeck, O. C. A re-investigation of the thio effect at the hammerhead cleavage site. *Nucleic Acids Res.* **27**, 479–484 (1999).
- Peracchi, A., Beigelman, L., Scott, E. C., Uhlenbeck, O. C. & Herschlag, D. Involvement of a specific metal ion in the transition of the hammerhead ribozyme to its catalytic conformation. *J. Biol. Chem.* **272**, 26822–26826 (1997).
- Wang, S., Karbstein, K., Peracchi, A., Beigelman, L. & Herschlag, D. Identification of the hammerhead ribozyme metal ion binding site responsible for rescue of the deleterious effect of a cleavage site phosphorothioate. *Biochemistry* **38**, 14363–14378 (1999).
- Murray, J. B. & Scott, W. G. Does a single metal ion bridge the A-9 and scissile phosphate groups in the catalytically active hammerhead ribozyme structure? *J. Mol. Biol.* **296**, 33–41 (2000).
- O'Rear, J. L. *et al.* Comparison of the hammerhead cleavage reactions stimulated by monovalent and divalent cations. *RNA* **7**, 537–545 (2001).
- Curtis, E. A. & Bartel, D. P. The hammerhead cleavage reaction in monovalent cations. *RNA* **7**, 546–552 (2001).
- Ferre-D'Amare, A. R., Zhou, K. & Doudna, J. A. Crystal structure of a hepatitis delta virus ribozyme. *Nature* **395**, 567–574 (1998).
- Rupert, P. B. & Ferre-D'Amare, A. R. Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature* **410**, 780–786 (2001).
- Rajagopal, P. & Feigon, J. Triple-strand formation in the homopurine:homopyrimidine DNA oligonucleotides d(G-A)4 and d(T-C)4. *Nature* **339**, 637–640 (1989).
- Sklenar, V. & Feigon, J. Formation of a stable triplex from a single DNA strand. *Nature* **345**, 836–838 (1990).
- Connell, G. J. & Yarus, M. RNAs with dual specificity and dual RNAs with similar specificity. *Science* **264**, 1137–1141 (1994).
- Legault, P. & Pardi, A. In situ probing of adenine protonation in RNA by ¹³C NMR. *J. Am. Chem. Soc.* **116**, 8390–8391 (1994).
- Ravindranathan, S., Butcher, S. E. & Feigon, J. Adenine protonation in domain B of the hairpin ribozyme. *Biochemistry* **39**, 16026–16032 (2000).
- Shih, I. H. & Been, M. D. Involvement of a cytosine side chain in proton transfer in the rate-determining step of ribozyme self-cleavage. *Proc. Natl Acad. Sci. USA* **98**, 1489–1494 (2001).
- Nakano, S., Chadalavada, D. M. & Bevilacqua, P. C. General acid-base catalysis in the mechanism of a hepatitis delta virus ribozyme. *Science* **287**, 1493–1497 (2000).
- Nakano, S. & Bevilacqua, P. C. Proton inventory of the genomic HDV ribozyme in Mg²⁺-containing solutions. *J. Am. Chem. Soc.* **123**, 11333–11334 (2001).
- Luptak, A., Ferre-D'Amare, A. R., Zhou, K., Zilm, K. W. & Doudna, J. A. Direct pK_a measurement of the active-site cytosine in a genomic hepatitis delta virus ribozyme. *J. Am. Chem. Soc.* **123**, 8447–8452 (2001).
- Nakano, S., Proctor, D. J. & Bevilacqua, P. C. Mechanistic characterization of the HDV genomic ribozyme: assessing the catalytic and structural contributions of divalent metal ions within a multichannel reaction mechanism. *Biochemistry* **40**, 12022–12038 (2001).
- Ryder, S. P. *et al.* Investigation of adenosine base ionization in the hairpin ribozyme by nucleotide analog interference mapping. *RNA* **7**, 1454–1463 (2001).
- Hampel, A. & Cowan, J. A. A unique mechanism for RNA catalysis: the role of metal cofactors in hairpin ribozyme cleavage. *Chem. Biol.* **4**, 513–517 (1997).
- Nesbitt, S., Hegg, L. A. & Fedor, M. J. An unusual pH-independent and metal-ion-independent mechanism for hairpin ribozyme catalysis. *Chem. Biol.* **4**, 619–630 (1997).
- Walter, N. G. & Burke, J. M. The hairpin ribozyme: structure, assembly and catalysis. *Curr. Opin. Chem. Biol.* **2**, 303 (1998).
- Cech, T. R. & Herschlag, D. (eds) *Group I Ribozymes: Substrate Recognition, Catalytic Strategies and Comparative Mechanistic Analysis* (Springer, Berlin, 1996).
- Narlikar, G. J. & Herschlag, D. Mechanistic aspects of enzymatic catalysis: lessons from comparison of RNA and protein enzymes. *Annu. Rev. Biochem.* **66**, 19–59 (1997).
- Shan, S., Kravchuk, A. V., Piccirilli, J. A. & Herschlag, D. Defining the catalytic metal ion interactions in the *Tetrahymena* ribozyme reaction. *Biochemistry* **40**, 5161–5171 (2001).
- Cate, J. H. *et al.* Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* **273**, 1678–1685 (1996).
- Juneau, K., Podell, E., Harrington, D. J. & Cech, T. R. Structural basis of the enhanced stability of a mutant ribozyme domain and a detailed view of RNA-solvent interactions. *Structure (Camb.)* **9**, 221–231 (2001).
- Golden, B. L., Gooding, A. R., Podell, E. R. & Cech, T. R. A preorganized active site in the crystal structure of the *Tetrahymena* ribozyme. *Science* **282**, 259–264 (1998).
- Michel, F. & Westhof, E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* **216**, 585–610 (1990).
- Szewczak, A. A. *et al.* An important base triple anchors the substrate helix recognition surface within the *Tetrahymena* ribozyme active site. *Proc. Natl Acad. Sci. USA* **96**, 11183–11188 (1999).
- Gordon, P. M., Sontheimer, E. J. & Piccirilli, J. A. Kinetic characterization of the second step of group II intron splicing: role of metal ions and the cleavage site 2'-OH in catalysis. *Biochemistry* **39**, 12939–12952 (2000).
- Sigel, R. K., Vaidya, A. & Pyle, A. M. Metal ion binding sites in a group II intron core. *Nature Struct. Biol.* **7**, 1111–1116 (2000).
- Gordon, P. M. & Piccirilli, J. A. Metal ion coordination by the AGC triad in domain 5 contributes to group II intron catalysis. *Nature Struct. Biol.* **8**, 893–898 (2001).
- Jestin, J. L., Deme, E. & Jacquier, A. Identification of structural elements critical for inter-domain interactions in a group II self-splicing intron. *EMBO J.* **16**, 2945–2954 (1997).
- Boudvillain, M., de Lencastre, A. & Pyle, A. M. A tertiary interaction that links active-site domains to the 5' splice site of a group II intron. *Nature* **406**, 315–318 (2000).
- Chu, V. T., Adamidi, C., Liu, Q., Perlman, P. S. & Pyle, A. M. Control of branch-site choice by a group II intron. *EMBO J.* **20**, 6866–6876 (2001).
- Zhang, L. & Doudna, J. A. Structural insights into group II intron catalysis and branch-site selection. *Science* **295**, 2084–2088 (2002).
- Costa, M., Michel, F. & Westhof, E. A three-dimensional perspective on exon binding by a group II self-splicing intron. *EMBO J.* **19**, 5007–5018 (2000).
- Swisher, J., Duarte, C. M., Su, L. J. & Pyle, A. M. Visualizing the solvent-inaccessible core of a group II intron ribozyme. *EMBO J.* **20**, 2051–2061 (2001).
- Yang, J., Zimmerly, S., Perlman, P. S. & Lambowitz, A. M. Efficient integration of an intron RNA into double-stranded DNA by reverse splicing. *Nature* **381**, 332–335 (1996).
- Frank, D. N. & Pace, N. R. Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu. Rev. Biochem.* **67**, 153–180 (1998).

61. Morl, M. & Marchfelder, A. The final cut. The importance of tRNA 3'-processing. *EMBO Rep.* **2**, 17–20 (2001).
62. Warnecke, J. M., Held, R., Busch, S. & Hartmann, R. K. Role of metal ions in the hydrolysis reaction catalyzed by RNase P RNA from *Bacillus subtilis*. *J. Mol. Biol.* **290**, 433–445 (1999).
63. Warnecke, J. M., Sontheimer, E. J., Piccirilli, J. A. & Hartmann, R. K. Active site constraints in the hydrolysis reaction catalyzed by bacterial RNase P: analysis of precursor tRNAs with a single 3'-S-phosphorothiolate internucleotide linkage. *Nucleic Acids Res.* **28**, 720–727 (2000).
64. Westhof, E. & Altman, S. Three-dimensional working model of M1 RNA, the catalytic RNA subunit of ribonuclease P from *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **91**, 5133–5137 (1994).
65. Harris, M. E., Kazantsev, A. V., Chen, J. L. & Pace, N. R. Analysis of the tertiary structure of the ribonuclease P ribozyme-substrate complex by site-specific photoaffinity crosslinking. *RNA* **3**, 561–576 (1997).
66. Frank, D. N., Adamidi, C., Ehringer, M. A., Pitulle, C. & Pace, N. R. Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA. *RNA* **6**, 1895–1904 (2000).
67. Li, Y. & Altman, S. A subunit of human nuclear RNase P has ATPase activity. *Proc. Natl Acad. Sci. USA* **98**, 441–444 (2001).
68. Xiao, S., Houser-Scott, F. & Engelke, D. R. Eukaryotic ribonuclease P: increased complexity to cope with the nuclear pre-tRNA pathway. *J. Cell. Physiol.* **187**, 11–20 (2001).
69. Tesmer, J. J. *et al.* Two-metal-ion catalysis in adenylate cyclase. *Science* **285**, 756–760 (1999).
70. Wyckoff, H. W. *et al.* The three-dimensional structure of ribonuclease-S. Interpretation of an electron density map at a nominal resolution of 2 Å. *J. Biol. Chem.* **245**, 305–328 (1970).
71. Drum, C. L. *et al.* Structural basis for the activation of anthrax adenylate cyclase exotoxin by calmodulin. *Nature* **415**, 396–402 (2002).
72. Treiber, D. K. & Williamson, J. R. Exposing the kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.* **9**, 339–345 (1999).
73. Thirumalai, D. & Woodson, S. A. Maximizing RNA folding rates: a balancing act. *RNA* **6**, 790–794 (2000).
74. Thirumalai, D., Lee, N., Woodson, S. A. & Klimov, D. Early events in RNA folding. *Annu. Rev. Phys. Chem.* **52**, 751–762 (2001).
75. Treiber, D. K. & Williamson, J. R. Beyond kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.* **11**, 309–314 (2001).
76. Zhuang, X. *et al.* A single-molecule study of RNA catalysis and folding. *Science* **288**, 2048–2051 (2000).
77. Liphardt, J., Onoa, B., Smith, S. B., Tinoco, I. J. & Bustamante, C. Reversible unfolding of single RNA molecules by mechanical force. *Science* **292**, 733–737 (2001).
78. Russell, R. *et al.* Exploring the folding landscape of a structured RNA. *Proc. Natl Acad. Sci. USA* **99**, 155–160 (2002).
79. Caprara, M. G., Mohr, G. & Lambowitz, A. M. A tyrosyl-tRNA synthetase protein induces tertiary folding of the group I intron catalytic core. *J. Mol. Biol.* **257**, 512–531 (1996).
80. Weeks, K. M. & Cech, T. R. Assembly of a ribonucleoprotein catalyst by tertiary structure capture. *Science* **271**, 345–348 (1996).
81. Chanfreau, G. & Jacquier, A. An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J.* **15**, 3466–3476 (1996).
82. Cohen, S. B. & Cech, T. R. Dynamics of thermal motions within a large catalytic RNA investigated by cross-linking with thiol-disulfide interchange. *J. Am. Chem. Soc.* **119**, 6259–6268 (1997).
83. Profenno, L. A., Kierzek, R., Testa, S. M. & Turner, D. H. Guanosine binds to the *Tetrahymena* ribozyme in more than one step, and its 2'-OH and the nonbridging pro-Sp phosphoryl oxygen at the cleavage site are required for productive docking. *Biochemistry* **36**, 12477–12485 (1997).
84. Murchie, A. I., Thomson, J. B., Walter, F. & Lilley, D. M. Folding of the hairpin ribozyme in its natural conformation achieves close physical proximity of the loops. *Mol. Cell* **1**, 873–881 (1998).
85. Andersen, A. A. & Collins, R. A. Rearrangement of a stable RNA secondary structure during VS ribozyme catalysis. *Mol. Cell* **5**, 469–478 (2000).
86. Noller, H. F., Hoffarth, V. & Zimniak, L. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **256**, 1416–1419 (1992).
87. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
88. Welch, M., Chastang, J. & Yarus, M. An inhibitor of ribosomal peptidyl transferase using transition-state analogy. *Biochemistry* **34**, 385–390 (1995).
89. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
90. Polacek, N., Gaynor, M., Yassin, A. & Mankin, A. S. Ribosomal peptidyl transferase can withstand mutations at the putative catalytic nucleotide. *Nature* **411**, 498–501 (2001).
91. Thompson, J. *et al.* Analysis of mutations at residues A2451 and G2447 of 23S rRNA in the peptidyltransferase active site of the 50S ribosomal subunit. *Proc. Natl Acad. Sci. USA* **98**, 9002–9007 (2001).
92. Murray, J. M. & Doudna, J. A. Creative catalysis: pieces of the RNA world jigsaw. *Trends Biochem. Sci.* **26**, 699–701 (2001).
93. Kumar, R. K. & Yarus, M. RNA-catalyzed amino acid activation. *Biochemistry* **40**, 6998–7004 (2001).
94. Illangasekare, M. & Yarus, M. Specific, rapid synthesis of Phe-RNA by RNA. *Proc. Natl Acad. Sci. USA* **96**, 5470–5475 (1999).
95. Illangasekare, M. & Yarus, M. A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. *RNA* **5**, 1482–1489 (1999).
96. Collins, C. A. & Guthrie, C. The question remains: is the spliceosome a ribozyme? *Nature Struct. Biol.* **7**, 850–854 (2000).
97. Yean, S. L., Wuenschell, G., Termini, J. & Lin, R. J. Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome. *Nature* **408**, 881–884 (2000).
98. Valadkhan, S. & Manley, J. L. Splicing-related catalysis by protein-free snRNAs. *Nature* **413**, 701–707 (2001).

Acknowledgements

We acknowledge D. Battle for extensive help with figure preparation, and V. Rath for comments on the manuscript.

The involvement of RNA in ribosome function

Peter B. Moore*† & Thomas A. Steitz*†‡

*Department of Molecular Biophysics and Biochemistry, and †Department of Chemistry, Yale University, PO Box 208107, New Haven, Connecticut 06520-8107, USA (e-mail: peter.moore@yale.edu)

‡Howard Hughes Medical Institute, New Haven, Connecticut 06520-8114, USA

The ribosome is a particle made of RNA and protein that is found in abundance in all cells that are actively making protein. It catalyses the messenger RNA-directed synthesis of proteins. Recent structural work has demonstrated a profound involvement of the ribosome's RNA component in all aspects of its function, supporting the hypothesis that proteins were added to the ribosome late in its evolution.

The discovery of the ribosome and the elucidation of its role in gene expression was one of the main achievements of molecular biology in the 1950s and '60s. Early investigations revealed that ribosomes invariably consist of a large and a small subunit, the former being roughly twice the molecular mass of the latter, and that both subunits are composed of ~60% RNA by weight. The small subunit mediates the interactions between messenger RNA (mRNA) and transfer RNAs (tRNAs) that determine the sequences of the proteins ribosomes make. The large subunit catalyses peptide bond formation¹.

That ribosomes contain protein surprised no one in the 1950s, especially after it became clear that they catalyse protein synthesis. Everyone knew that enzymes are proteins. The surprise was that ribosomes contain RNA. At the time, three hypotheses could be entertained about the contribution RNA makes to ribosome function: (1) it is the substance that determines the sequences of the proteins ribosomes make; (2) it is the (inert) structural scaffold on which the catalytically active, proteinaceous parts of the protein synthetic apparatus are assembled; or (3) it contributes directly to the catalytic processes of protein synthesis. The discovery of mRNA around 1960 put an end to the first possibility, and left the field in a quandary. The second hypothesis was unappealing because it does not make evolutionary sense^{2,3}. If protein performs the key functions of the ribosome now, it probably did so when the ribosome first evolved, and if that is so, why does the ribosome contain RNA? Furthermore, if the first proteins were synthesized by proteins in the ribosome, what made the first ribosomal proteins? The third alternative was troubling because it challenged the doctrine that enzymes are proteins.

With the progression of time, the third hypothesis has gained increasing support. The idea that ribosomal RNAs (rRNAs) might participate directly in protein synthesis became entirely credible in the early 1980s, following the discovery that RNAs can indeed catalyse chemical reactions^{4,5}. In addition, genetic and biochemical evidence slowly accumulated that implicated rRNA in ribosome function⁶⁻⁸. Nevertheless, a full appreciation of the functional importance of rRNA in the ribosome has emerged only recently. The more than two-dozen crystal structures of ribosomes and their ligand complexes now available document the involvement of RNA in ribosome activity at a level that even the most ardent advocates of the functional

importance of rRNA could hardly have anticipated. It is the proteins in the ribosome that perform a largely structural function, not the RNA.

Ribosome active sites

The functional importance of RNA in the ribosome, relative to protein, can best be evaluated by examining its prominence in the small fraction of the ribosome's total mass found in its active sites, rather than by measuring what it contributes to the particle as a whole. That said, it is almost certainly true that the overall abundance of RNA in the modern ribosome reflects an evolutionary history that started with an all-RNA particle. Presumably, this original proto-ribosome had decoding and peptide synthesis sites that were similar to the ones we see today, and proteins were added to it later to enhance its function. But before such issues can be approached intelligently, it is important to consider what has been learned from biochemical studies about the

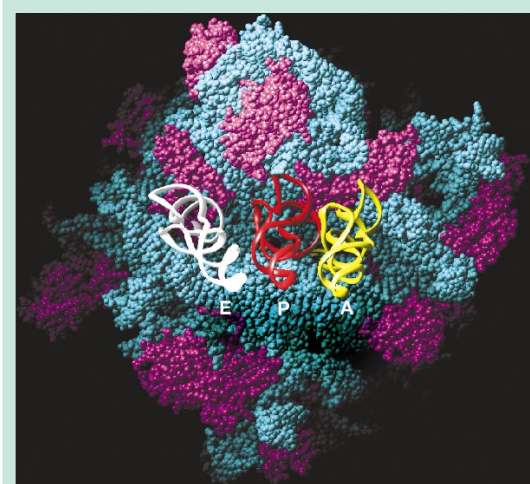


Figure 1 The arrangement of tRNA-binding sites on the large ribosomal subunit. The surface of the *H. marismortui* large ribosomal subunit that interacts with the small subunit is shown with its RNA portions in blue, and its protein regions in purple. tRNAs are shown in ribbon format bound to the E, P and A sites. The positions they occupy are deduced from the 70S structure of Noller and colleagues³⁹. (Reproduced with permission from ref. 14.)

active sites of the eubacterial ribosome. Archaeal and eukaryotic ribosomes work in much the same way.

Aminoacyl-tRNAs are the substrates that ribosomes consume during protein synthesis. Every cell contains a population of tRNAs that differ in sequence, but have similar relative molecular mass (~25,000), similar secondary structures, and the same L-shaped tertiary structure. At the distal end of the longer arm of the L there is a three-base, anticodon sequence in every tRNA that is complementary to one of the mRNA base triplets that encodes a specific amino acid. At the distal end of the short arm of the L of every tRNA is a 3'-terminal CCA sequence to which the amino acid specified by the anticodon is attached⁹. During protein synthesis, the anticodon ends of tRNAs interact with the mRNA bound to the small subunit and their aminoacylated CCA sequences interact with the large subunit.

Amino acids are attached to tRNAs by enzymes called aminoacyl-tRNA synthetases, of which there is one for each kind of amino acid. They catalyse the formation of ester bonds between the α -carboxylate groups of amino acids and the 2'- or 3'-hydroxyl groups of the 3'-terminal adenine residue of tRNAs. Synthetases translate the genetic code with great specificity by attaching one type of amino acid to all tRNAs bearing the appropriate anticodons. There is at least one kind of tRNA molecule for every amino acid used in protein synthesis.

Ribosomes assemble proteins one amino acid at a time, starting with the N-terminal residue¹⁰. The reaction that results in amino acid polymerization is the nucleophilic attack of the α -amino group of an amino acid esterified to one tRNA on the carbonyl carbon of the ester linking a peptide (or amino acid) to a second tRNA. Resolution of the resulting tetrahedral intermediate deacylates the tRNA that is the carbonyl group donor, and leaves the tRNA that provided the α -amino group esterified to a peptide that has been extended by one amino acid. This reaction occurs at the peptidyl transferase centre of the large ribosomal subunit, which includes a sub-site to which the CCA-peptide moiety of peptidyl-tRNAs binds (the P site), and a sub-site that interacts with the amino acid-carrying CCA end of aminoacyl-tRNA (the A site).

Before another amino acid can be added a growing peptide chain, the deacylated tRNA in the P site must be replaced by the peptidyl-

tRNA resident in the A site, and a new aminoacyl-tRNA must be placed in the A site. The exchange of P-site tRNAs required results from a still poorly understood conformational transformation called translocation, which involves both subunits (see below). It is mediated by a G-protein called elongation factor G (EF-G) in prokaryotes, and it is accompanied by the cleavage of guanosine triphosphate (GTP).

Delivery of the correct aminoacyl-tRNA to the ribosome is the sequence-determining step of protein synthesis. If the polypeptide being made by a ribosome is to have the sequence required by the mRNA bound to it, only one of the many different aminoacyl-tRNAs present in the cell can be accepted during each cycle of peptide chain elongation. The selection required is accomplished by the small subunit's decoding centre. Aminoacyl-tRNAs are delivered to the ribosome bound to a second G-protein factor, elongation factor Tu (EF-Tu) and GTP. These so-called ternary complexes bind tightly to the ribosome only if the anticodon sequences of their tRNA components are complementary to the mRNA codon presented in the decoding centre's A sub-site. If the match is satisfactory, acceptance occurs, a process that involves cleavage of GTP, followed the release of EF-Tu-GDP from the ribosome, and a major change in the orientation of the tRNA on the ribosome.

The decoding centre also has a P sub-site. The codon in the small subunit's P site interacts with the anticodon of a peptidyl-tRNA bound in the large subunit's P site at the time that peptide bond formation occurs. During translocation, the anticodon end of a tRNA in the A site moves into the small subunit's P site, displacing the deacylated tRNA already bound there (if any), and the mRNA bound to the small subunit advances in sympathy in the 5' direction by three nucleotides. This results in the presentation of a new codon in the A site so that the next cycle of aminoacyl-tRNA selection, peptide bond formation and translocation can occur.

In addition to its peptidyl transferase and decoding centres, the ribosome includes a factor-binding centre. It is part of the large subunit, and all of the G-protein factors involved in protein synthesis interact with it during at least part of their duty cycles. In some poorly understood way, it triggers the GTPase activities of these proteins, and mediates the conformational changes they all facilitate. That

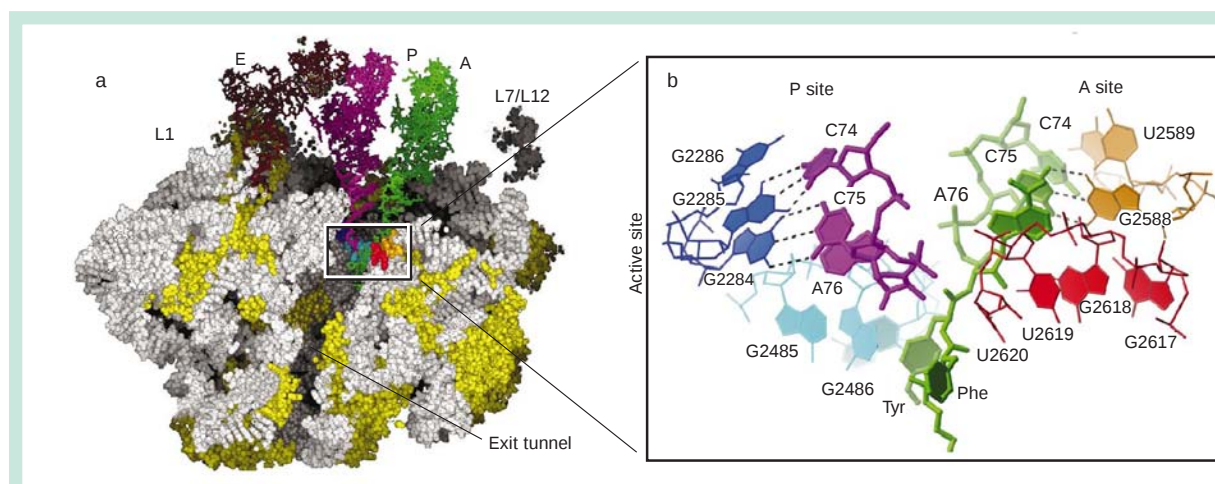


Figure 2 Interactions of the CCA ends of ribosome-bound tRNA with the large ribosomal subunit. **a**, Cut-away view of the large ribosomal subunit with tRNAs bound. tRNAs are positioned on the large ribosomal subunit as described in the legend for Fig. 1, and the subunit sliced in half along a plane approximately perpendicular to the Fig. 1 plane of view to reveal the placement of the acceptor stems of ribosome-bound tRNAs in the large subunit's peptide exit tunnel. rRNA is shown in white, and ribosomal protein in yellow. tRNAs are colour coded as follows: E site, brown; P site, purple; and A site, green. **b**, Interactions of the CCA sequences bound in the P site and the A site with 23S

rRNA. The molecule bound in the P site is the deacylated tRNA analogue CCA (purple). The molecule bound in the A site is the peptidyl-tRNA analogue CCA-puromycin-phenylalanine-caproic acid-biotin (green)¹⁵. Blue bases are components of the P loop²⁹; brown bases belong to the A loop³⁰. The nucleotides in cyan and red are other components of the peptidyl transferase centre. Bases are numbered according to the sequence of *H. marismortui* 23S rRNA. The two 23S rRNA bases closest to the newly formed peptide bond are A2486 and U2620, which correspond to A2451 and U2585 in *E. coli*, respectively. (Reproduced with permission from ref. 15.)

said, it is important to realize that these factors act catalytically to enhance properties the ribosomal system already possesses. Ribosomes programmed with mRNAs will select aminoacyl-tRNAs correctly in the absence of EF-Tu. Furthermore, the peptidyl transferase reaction occurs spontaneously on any ribosome that has appropriate substrates bound to it. Even translocation can occur in the absence of EF-G. In fact, ribosomes programmed with mRNAs of appropriate sequence will slowly synthesize oligopeptides in the absence of factors¹¹, although in the presence of factors protein synthesis proceeds much faster than it otherwise would, and the accuracy with which mRNAs are translated increases.

Even though the chemical logic of protein synthesis would seem to require the existence of no more than two tRNA-binding sites on the ribosome, there is a third, the E site. During translocation deacylated tRNAs bound to the P site move into the E site¹². Release of E site-bound tRNAs into solution accompanies the acceptance of the next aminoacyl-tRNA in the A site, not the translocation step of the next cycle of elongation. Like the A and P sites, the E site has components on both subunits that adjoin the decoding centre and the peptidyl transferase centre¹³.

The peptidyl transferase centre

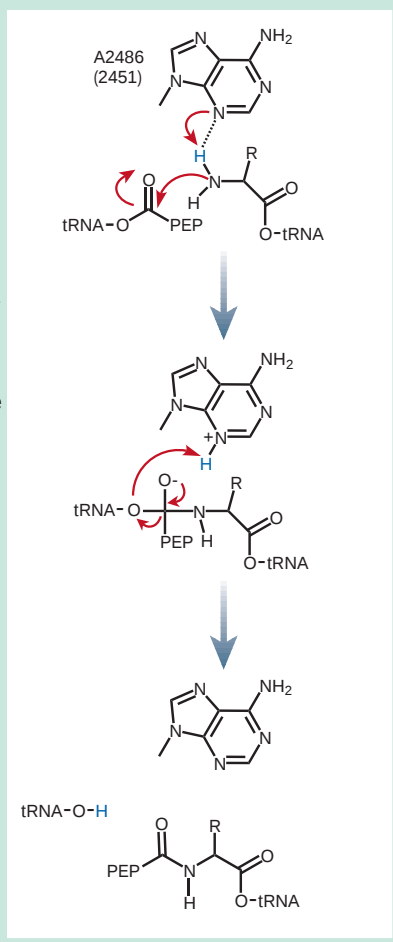
Much of what is known today about the structures and mechanisms of action of the decoding and peptidyl transferase centres has been deduced from crystals of isolated subunits. Ideally, this information would have been obtained from crystals of 70S ribosomes with appropriate ligands bound, but there are currently no crystals of the 70S ribosome that diffract to the resolution required for independent determination of atomic structure. The reason subunit crystals are useful in this context is that the peptidyl transferase and decoding centres are contained entirely within single, separated subunits; thus, activities that are obviously related to those that the two centres display in 70S ribosomes can be elicited from isolated subunits. The data obtained must be interpreted carefully, however, as the low resolution of the available 70S electron density maps means that small differences may exist between the structures of these centres in isolated subunits and in 70S ribosomes.

The crystal structures of the large subunit bound with substrates, substrate analogues and products show that the peptidyl transferase centre is at the bottom of a large cleft in its small-subunit-binding surface^{14,15}. Figure 1 shows that face of the large subunit with tRNAs placed in the E, P and A sites; their acceptor stems disappear into the cleft. As Fig. 2a reveals, the terminal CCA sequences of the A- and P-site-bound tRNAs, which must interact during peptide bond formation, meet at the small-subunit end of a tunnel that passes through the subunit to its back side. (Fig. 2a is derived from Fig. 1 by rotating the top of the subunit ~90° away from the viewer, and then slicing the subunit vertically along a plane that includes the acceptor stems of the bound tRNAs and the axis of the tunnel; the 'hemisphere' of the subunit closest to the viewer has been removed, and the tRNAs left in place.) The two CCA sequences meet at the site where peptidyl transferase substrate analogues and products bind to the subunit (Fig. 2b).

None of these discoveries was a surprise. Electron microscopic evidence for the tunnel (appropriately named the 'peptide exit tunnel') first emerged in the 1980s^{16,17}, following the demonstration that nascent, ribosome-bound peptides first become accessible to solvent on the back side of the large ribosome^{18,19}. All doubts about its existence were resolved by the cryoelectron microscopic studies reported in the late 1990s that also demonstrated that the peptidyl transferase centre is located at its small subunit end^{20–22}.

The peptidyl transferase centre is formed by nucleotides from domain V of the principal RNA component of the large subunit, 23S rRNA; most of these nucleotides are components of its central loop, consistent with earlier biochemical results^{23,24}. In addition to verifying these biochemical results, these crystal structures prove that there is no protein whatsoever in the peptidyl transferase centre of the ribosome. Thus the most fundamental of the ribosome's activities, its

Figure 3 A possible mechanism for the involvement of A2486 (*H. marismortui*)/A2451 (*E. coli*) in peptide bond formation. One of the hydrogens belonging to the α -amino group of the aminoacyl-tRNA bound in the A site of the large subunit seems to interact with the N3 of A2451 (*E. coli* numbering) during the nucleophilic attack of that amino group on the carbonyl carbon of the ester linking the nascent peptide to the tRNA in the P site (top frame). If that interaction were strong enough, it could enhance the rate of peptide bond formation by abstracting a proton from the attacking α -amino group (middle and bottom frames).



peptide bond-forming activity, is catalysed by a structure composed entirely of RNA¹⁴.

Precise substrate alignment is an important source of the catalytic power in the peptidyl transferase centre¹⁴, as indeed it is in all enzymes^{25–27}. It had been proposed earlier on theoretical grounds that substrate orientation might be important in the ribosome²⁸, and much of the alignment required is ensured by base-pairing interactions between the CCA sequences of P-site- and A-site-bound tRNA with nucleotides in the so-called P loop (helix 80 of 23S rRNA)²⁹ and A loop (helix 92)³⁰, respectively (Fig. 2b). These interactions position the α -amino group of an aminoacyl-tRNA in the A site so that it can attack the carbonyl carbon of the ester linking a polypeptide to a tRNA bound in the P site¹⁴.

Structures of the large ribosomal subunit of the halophilic archaeon *Haloarcula marismortui* with substrate and transition-state analogues bound suggest that the peptidyl transferase centre may derive catalytic power from a second source. The N3 of residue A2451 (*Escherichia coli* numbering) of 23S rRNA is hydrogen bonded to the α -amino group of aminoacyl-tRNAs in the A site¹⁴. Not only does this interaction contribute to the positioning of that group, it could further enhance the rate of peptide bond formation by helping abstract a proton from that α -amino group (Fig. 3).

Biochemical data obtained recently using ribosomes mutated at position 2451 suggest that A2451 may indeed act as a general base during peptide bond formation. First, all such mutations are dominant lethal in *E. coli*, consistent with the view that A2451 is critical to ribosome function^{31–33}. Second, recent kinetic studies show that the rate of the chemical step of peptide bond formation increases by a factor of 100–150 when a group in 70S ribosomes that has a pK_a near 7.5 is deprotonated, and that this titration effect is not seen in ribosomes

that have uracil at position 2451 instead of adenine³⁴. These observations contradict conclusions published previously that were also based on measurements of the pH dependence of the rate of peptide bond formation, but under the conditions used in these earlier experiments, the rate of the chemical step of the reaction seems not have been rate limiting^{32,35}.

The simplest interpretation of these observations may be the right one; A2451 may be the group whose titration affects the rate of peptide bond formation, and it may do so because it functions as a general base during peptide bond formation. However, it would be premature to conclude that the issue is settled. It has not been proven that it is the titration of A2451 that affects the rate of peptide bond formation, only that the nucleotide at position 2451 must be an adenine for the titration effect to be seen. The pK_a of the N3 of A2451 would have to be 7.5 both to function as a general base under physiological conditions, as proposed¹⁴, and to explain the titration data. But the pK_a of the N3 of an unperturbed adenosine is only about 1.0. At the time the A2451 hypothesis was advanced, chemical data existed that seemed to show that the pK_a of A2451 is unusually high³¹, but it is now clear that those data are irrelevant to the properties of A2451 in ribosomes competent in peptide bond formation^{33,36,37}. Finally, mutation of one of the bases postulated earlier to be crucial for perturbing the pK_a of A2451 (ref. 14) does not affect viability and is also reported not to affect the rate of peptide bond formation^{32,33}. For all these reasons, it remains possible that a pH-dependent conformational change of some kind that depends on the identity of the nucleotide at position 2451 affects the rate of peptide bond synthesis.

The decoding centre

A remarkable amount of information about the decoding centre has been generated from crystal structures of the small subunit, in part as a result of a fortuitous accident³⁸. The small subunit of *Thermus thermophilus* ribosomes has a protruding stem-loop structure, called the spur, whose conformation resembles that of the anticodon stem-loop of a tRNA. In addition, the 3' end of 16S rRNA folds back into the P site, where it binds as though it were a piece of mRNA. One of the intermolecular interactions that stabilizes crystals of these subunits involves the insertion of the spur of one subunit into the P site of a neighbour. The resulting interaction is similar to that of a tRNA bound to the P site of 70S ribosomes with its small subunit component^{39,40}, and thus much can be learned about the tRNA/P-site interaction from crystals of small subunits that contain no tRNA or mRNA.

Unlike the peptidyl transferase centre, the decoding centre includes protein. The minor groove of the anticodon stem of a tRNA bound to the P site contacts not only nucleotides belonging to the 3' major domain and the central domain of the RNA component of the small subunit, 16S rRNA, but also amino acids belonging to ribosomal proteins S13 and S9. The anticodon of tRNA bound to the P site interacts with the mRNA codon exposed at the bottom of that site as well as the penultimate helix (helix 44). The A site is also a 'hybrid' structure^{38,41}. It includes residues from ribosomal protein S13, and a loop of S12 abuts the mRNA codon exposed there. The elements of 16S rRNA represented in the A site include the 530 loop, helix 44 and the 3' major domain. But despite these interactions between proteins and tRNA, the decoding centre, like the peptidyl transferase centre, is basically an RNA machine. One gets the impression that if the proteins in the decoding site could be removed without otherwise altering its structure, it would still function properly.

The decoding centre is located in the region where the head of the small subunit meets its body, and it includes the most important parts of the site that interacts with mRNA. Parts of that site have been mapped out in small subunit crystals that have mRNA analogues bound⁴¹, and the rest has been observed at lower resolution in 70S crystals carrying short, natural mRNA sequences⁴². A double helix forms between the Shine–Dalgarno sequence of appropriately positioned mRNAs and the complementary sequence at the 3' terminus

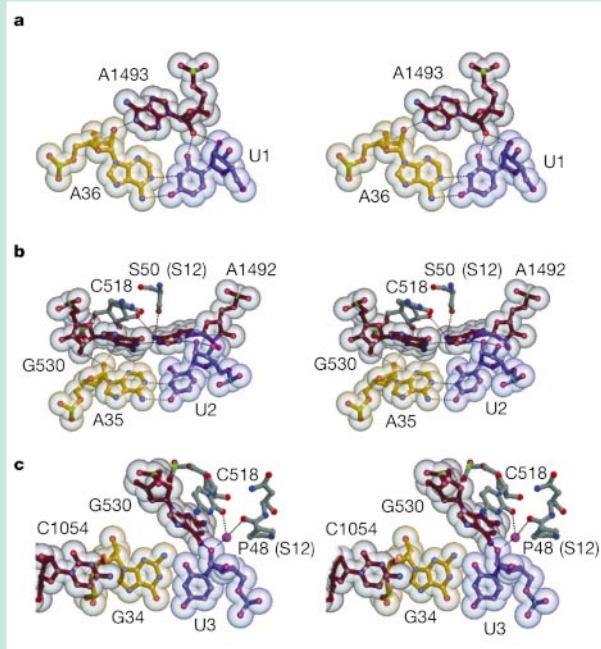


Figure 4 Fidelity-checking interactions in the A site of the small ribosomal subunit. This figure shows the interactions between ribosome residues and the base pairs that form when an anticodon stem-loop with a GAA anticodon interacts in the A site with a (cognate) UUU codon. **a**, The type I A-minor interaction between A1493 and the AU pair that forms between the U at the first position in the codon and the 3' A of the anticodon. **b**, The type II A-minor interaction that forms between the AU pair involving the second U in the mRNA codon and A1492, and the similar interaction with the AU pair involving G530, which also hydrogen bonds with A1492. **c**, The interaction of the ribosome with the so-called wobble-position base pair, which involves the 3' U of the mRNA codon. Note these interactions monitor only the placement of the mRNA base, not the placement of the anticodon base that is paired with it. (Reproduced from ref. 41 with permission.)

of 16S rRNA, as expected^{43,44}. The Shine–Dalgarno helix is found in the region between the subunit's head and platform, close to its E site. The P-site portion of bound mRNAs is hard to visualize in difference maps owing to the tendency of the 3' end of 16S rRNA to occupy that region in vacant ribosomes. Nevertheless, it is clear that there is a sharp kink in the backbone of mRNAs between codons in the P site and codons in the A site. On the 3' side of the A site, the molecule passes through a tunnel formed in part by proteins S3, S4 and S5, which may function as a helicase to remove secondary structure from mRNA as it enters the decoding region⁴².

The first step in the mechanism by which the small ribosomal subunit correctly decodes mRNAs seems to be accomplished entirely by 16S rRNA, and the conformation of the complex that forms between mRNA and tRNA in the A site depends on whether or not the interaction between codon and anticodon is cognate⁴¹. When a cognate tRNA enters the A site, the bases of A1492 and A1493 change positions and form type II and type I A-minor interactions⁴⁵, respectively, with the minor groove edge of the first two base pairs of the resulting codon–anticodon helix (Fig. 4). These interactions are not possible if the bases in the second and third positions of the anticodon of the tRNA in the A site do not form Watson–Crick base pairs with the first and second bases of the mRNA codon presented there. G530 responds to the codon–anticodon pairing that occurs with both the second and third bases in the mRNA codon. However, the interactions of 16S rRNA with the third base pair, which involve primarily G530, are less sensitive to base-pair geometry than the interactions it makes with the first two pairs. Thus these interactions not only

explain why the code is degenerate in the third position, but also why the ribosome is able to discriminate as well as it does between cognate tRNAs and near-cognate tRNAs. They are also likely to be important for the ribosome's capacity to enhance translational fidelity by proofreading.

Electron microscopic data show that the orientation of aminoacyl-tRNAs on the ribosome changes dramatically between the time they are delivered to the ribosome by EF-Tu and the time when peptide bond formation occurs^{13,46,47}. At the time of delivery, when tRNAs are initially selected, the anticodons of tRNAs are in the A site of the decoding centre, but their aminoacylated CCA sequences are nowhere near the peptidyl transferase centre. Accommodation — the reorientation that puts the CCA end of a tRNA into the A site of the peptidyl transferase centre — does not occur unless the tRNA delivered is cognate or near cognate to the mRNA sequence in the A site. The structures of small subunits with stem-loops bound and 70S ribosomes with tRNAs bound correspond to the post-accommodation state, and it is unclear how the geometry of the A-site interaction with tRNA anticodons in the post-accommodation state differs from that of the pre-accommodation state, whose interactions determine initially whether a tRNA is accepted or not.

Nevertheless, when a cognate interaction occurs in the A site, by a mechanism that is unknown, a signal is transmitted from the A site to the G domain of EF-Tu, which is bound to the factor-binding centre on the large subunit. That signal, which may involve the conformational change that occurs in the A site of the small subunit when a cognate codon-anticodon interaction occurs, triggers GTP hydrolysis, factor release from the ribosome, and the other steps of accommodation. These inter-subunit interactions will not be understood until high-resolution crystal structures are obtained of ribosomes complexed with factors and tRNAs trapped on the pre- and post-accommodation states.

The E site

The E site is richer in protein than the A and P sites. In addition to the several contacts that E-site-bound tRNAs make with both 16S and 23S rRNA, their anticodon stems interact extensively with S7 in the small subunit, and their T loops and T stems contact L1 in the large subunit. There are additional interactions involving L33 (ref. 40), and a substantial body of biochemical evidence exists indicating that the E site interacts allosterically with the A site¹².

The factor-binding centre

The GTPase domains of EF-Tu, EF-G and all the other G-protein factors involved in protein synthesis interact with the factor-binding centre during protein synthesis. Because there are no high-resolution structures of ribosomes with G-protein factors bound, our understanding of the centre's location and the way it interacts with factors is limited to what can be surmised from lower-resolution electron microscopic images and molecular biological investigations^{45–50}, combined with model building using atomic structures⁴⁴. The sarcin-ricin loop (stem-loop 95 of 23S rRNA) is one of the centre's critical components⁴⁸, but the centre also includes several proteins (for example, L7/12, L11, L6, L14; refs 49, 50), some of which are essential for its operation. The factors that interact with the centre undergo significant conformational changes as they perform their functions, as well as cleaving GTP, and the geometries of the complexes they form with the ribosome change in sympathy, as does the conformation of the ribosome itself^{51–55}.

Other ribosomal sites

The protein synthesis factors that are not G proteins bind to sites on the ribosome outside the factor-binding centre. High-resolution structures have been reported for two such factors bound to the small subunit: initiation factor 1 (IF1)⁵⁶, and the C-terminal domain of initiation factor 3 (IF3)⁵⁷. The site to which IF1 binds includes the A site of the small subunit's decoding centre, and its interaction with that

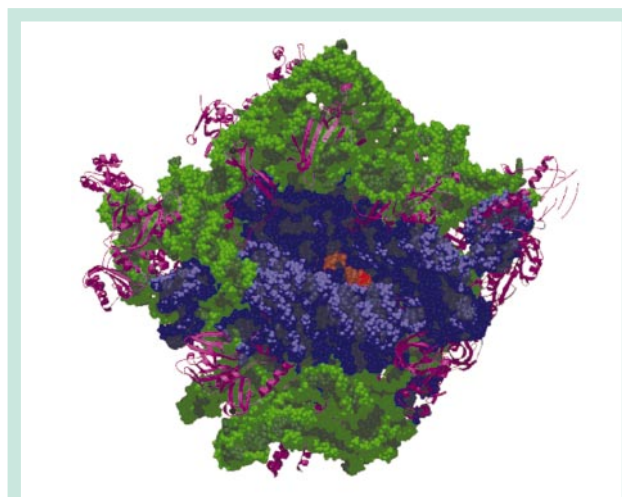


Figure 5 Conservation in the large ribosomal subunit. The large ribosomal subunit of *H. marismortui* is viewed from its small subunit interface side, looking down into its active site, where an inhibitor (red) is shown bound. RNA atoms are shown as space-filling spheres, and the protein is shown in ribbon format. The blue-coloured RNA is RNA belonging to the core of 23S/28S rRNA that is conserved in all species. The green-coloured RNA represents sequences that are outside the conserved core. (Reproduced with permission from *Science*.)

site causes conformational changes in the small subunit that interfere with its interaction with the large subunit, as well as with tRNAs. The C-terminal domain of IF3 binds to the solvent side of the platform in small subunit crystals into which it has been soaked, but as inter-subunit contacts in the crystals examined obstruct the IF3-binding site identified by others^{58,59}, the relevance of this observation is unclear.

There are now about 20 structures available for antibiotics bound to ribosomal subunits, and the number is increasing rapidly. These structures affirm the importance of RNA in ribosome function. Most of the antibiotics examined structurally have been found to inhibit ribosome function by binding to sites composed entirely of RNA. Many antibiotics that target the small subunit interfere with protein synthesis by inhibiting one or more of the conformational changes that accompany the normal function of that subunit^{38,41,57}. In contrast, antibiotics examined so far that inhibit the function of the large subunit antibiotics act in a different way. Their interactions with the ribosome sterically block the access of substrates to the peptidyl transferase centre or the passage of nascent peptides down the exit tunnel (refs 60, 61, and J. L. Hansen, N. Ban, P. Nissen, P.B.M. and T.A.S., in preparation).

Another functionally vital region of the ribosome that consists primarily of RNA is the interface between the two subunits in the 70S ribosome. It includes several sites, called bridges, where components of the two subunits contact each other. Information about these bridges can be gleaned only from structures of 70S ribosomes, and hence the resolution at which they are understood is currently limited. Nevertheless, it is clear that most bridges result from the interaction of 16S rRNA sequences with 23S rRNA sequences⁴⁰, and so are rich in RNA. It is also evident that the motions of tRNAs through the inter-subunit gap of the 70S ribosome that are essential for protein synthesis must be accompanied by the making and breaking of inter-subunit bridges. These structures play a dynamic role in protein synthesis.

The final ribosomal sites to consider are the region on the back side of the large ribosomal subunit that binds the translocon, which is the apparatus responsible for protein secretion, and the site that interacts with the signal recognition particle, which is responsible for the translational arrest that occurs early in the synthesis of secreted proteins so that ribosomes can interact with the translocon before

synthesis is completed. The structure of eukaryotic ribosomes bearing nascent polypeptides bound to Sec61 — the protein complex that forms the interface between secreting ribosomes and the rest of the translocon — has been investigated extensively by electron microscopy⁶². Sec61 is a more-or-less toroidal protein assembly that binds with its pore concentric with the distal end of the polypeptide exit tunnel. The contacts formed involve both elements of 23S/28S rRNA and the proteins that surround the end of the exit tunnel, L19, L23, L24 and L29 (ref. 62).

Evolution of the ribosome's active sites

It has often been emphasized that no enzyme as complex as the modern ribosome could possibly have emerged from the primordial soup in a single step. The first ribosomes were very likely composed entirely of RNA. The evidence for this hypothesis is that the functional core of the modern ribosome, its decoding site and its peptidyl transferase centre, consists primarily of RNA, and the bulk of its proteins are found on its surface, well removed from its functional centres (Fig. 5). The first ribosome was almost certainly a smaller, simpler structure than its modern descendants, and had a more limited function. The 'proto-ribosome' might have been a one-subunit object that catalysed peptide bond formation in an uncoded, non-processive way, and thus made small oligopeptides of random sequence that were not proteins in the modern sense, and may have had some other purpose⁶³. It probably gained its decoding function, and the small subunit with which that function is associated, later, in parallel with the evolution of the genome and of tRNA. Because the protein factors that facilitate ribosome function are still not absolutely essential, it is likely that they came later. The E site also seems to be a late evolutionary addition; one can easily imagine that the primitive ribosome discharged deacylated tRNAs directly from the P site into solution rather than retaining them bound.

The chemical compositions of the ribosome's functional centres correlate crudely with their evolutionary age in this scheme — the more recent the centre, the more protein it contains. The peptidyl transferase centre, the most ancient of them all, has no protein in it whatsoever. The decoding centre, the second to be added to the particle, is predominantly RNA, but protein does make a modest contribution to its function. No atomic-resolution crystal structures have been established yet for elongation factor/ribosome complexes, but both biochemical data and lower-resolution structural studies indicate not only that the factor-binding site contains several proteins, but also that many of them are critical for its function. Consistent with this pattern, the E site, which is a late addition, is protein rich.

Concluding comment

The importance of RNA in ribosome function has been proven beyond question by the crystal structures published in the past two years. But illuminating as these structures are, they provide only tantalizing hints about how the nucleotides in the ribosome's active centres perform their functions. A fertile area of inquiry has opened up that should challenge biochemists and molecular biologists for years to come.

A full understanding of ribosome function is unlikely to emerge without additional help from the structural biology community. Large conformational changes occur in the ribosome during protein synthesis, and in the absence of high-resolution structures for the 70S ribosome in many more states than have been characterized so far, it will be difficult to understand what these changes are and why they occur. Although it will not be easy to prepare the crystals required, the rewards are certain to be enormous. □

1. Tissieres, A. in *Ribosomes* (eds Nomura, M., Tissieres, A. & Lengyel, P.) 3–12 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1974).
2. Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
3. Woese, C. R. in *Ribosomes. Structure, Function, and Genetics* (eds Chambliss, G. et al.) 357–376 (University Park Press, Baltimore, 1980).
4. Cech, T., Zaugg, A. & Grabowski, P. *In vitro* slicing of the ribosomal RNA precursor of *Tetrahymena*: involvement of the guanosine nucleotide in the excision of the intervening sequence. *Cell* **27**,

- 487–496 (1981).
5. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytically active subunit of the enzyme. *Cell* **35**, 849–857 (1983).
6. Noller, H. F. Ribosomal RNA and translation. *Annu. Rev. Biochem.* **60**, 191–227 (1991).
7. Moore, P. B. in *The RNA World* (eds Gesteland, R. F. & Atkins, J. F.) 119–136 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1993).
8. Santer, M. & Dahlberg, A. E. in *Ribosomal RNA, Structure, Evolution, Processing, and Function in Protein Biosynthesis* (eds Zimmermann, R. A. & Dahlberg, A. E.) 3–20 (CRC Press, Boca Raton, 1995).
9. Soll, D. & RajBhandary, U. L. (eds) *tRNA. Structure, Biosynthesis, and Function* (American Society for Microbiology Press, Washington DC, 1995).
10. Green, R. & Noller, H. F. Ribosomes and translation. *Annu. Rev. Biochem.* **66**, 679–716 (1997).
11. Gavrilova, L. P., Kostashkina, O. E., Kotliansky, V. E., Rutkevitch, N. M. & Spirin, A. S. Factor-free ("non-enzymatic") and factor-dependent systems of translation of polyuridylic acid by *Escherichia coli* ribosomes. *J. Mol. Biol.* **101**, 537–552 (1976).
12. Rheinberger, H.-J. et al. in *The Ribosome. Structure, Function & Genetics* (eds Hill, W. E. et al.) 318–330 (ASM Press, Washington DC, 1990).
13. Agrawal, R. K. et al. Direct visualization of A-, P-, and E-site transfer RNAs in the *Escherichia coli* ribosome. *Science* **271**, 1000–1002 (1996).
14. Nissen, P., Ban, N., Hansen, J., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
15. Schmeing, T. M. et al. A pre-translocational intermediate in protein synthesis observed in crystals of enzymatically active 50S subunits. *Nature Struct. Biol.* **9**, 225–230 (2002).
16. Milligan, R. A. & Unwin, P. N. T. Location of the exit channel for nascent protein in 80S ribosome. *Nature* **319**, 693–695 (1986).
17. Yonath, A., Leonard, K. R. & Wittmann, H. G. A tunnel in the large ribosomal subunit revealed by three-dimensional image reconstruction. *Science* **236**, 813–816 (1987).
18. Bernabeu, C. & Lake, J. A. Nascent polypeptide chains emerge from the exit domain of the large ribosomal subunit: immune mapping of the nascent chain. *Proc. Natl Acad. Sci. USA* **79**, 3111–3115 (1982).
19. Bernabeu, C., Tobin, E. M., Fowler, A., Zabin, I. & Lake, J. A. Nascent polypeptide chains exit the ribosome in the same relative position in both eucaryotes and procaryotes. *J. Cell Biol.* **96**, 1471–1474 (1983).
20. Frank, J. et al. A model for protein synthesis based on cryo-electron microscopy of the *E. coli* ribosome. *Nature* **376**, 441–444 (1995).
21. Stark, H. et al. The 70S *Escherichia coli* ribosome at 23 Å resolution: fitting the ribosomal RNA. *Structure* **3**, 815–821 (1995).
22. Beckmann, R. et al. Alignment of conduits for the nascent polypeptide chain in the ribosome-Sec61 complex. *Science* **278**, 2123–2128 (1997).
23. Noller, H. F. Structure of ribosomal RNA. *Annu. Rev. Biochem.* **53**, 119–162 (1984).
24. Garrett, R. A. & Rodriguez-Fonseca, C. in *Ribosomal RNA. Structure, Evolution, Processing and Function in Protein Biosynthesis* (eds Zimmermann, R. A. & Dahlberg, A. E.) 327–355 (CRC Press, Boca Raton, 1996).
25. Koshland, D. E., Caraway, K. W., Dafforn, G. A., Gass, J. D. & Storm, D. R. The importance of orientation factors in enzymatic reactions. *Cold Spring Harb. Symp. Quant. Biol.* **36**, 13–20 (1971).
26. Koshland, D. E. Molecular basis of enzyme catalysis and control. *Pure Appl. Chem.* **25**, 119– (1971).
27. Page, M. I. & Jencks, W. P. Aminolysis of acetylhydrazide and rate acceleration caused by intramolecular catalysis. *Fed. Proc.* **30**, 1240 (1971).
28. Nierhaus, K. H., Schulze, H. & Cooperman, B. S. Molecular mechanisms of the ribosomal peptidyl transferase center. *Biochem. Int.* **1**, 185–192 (1980).
29. Samaha, R. R., Green, R. & Noller, H. F. A base pair between tRNA and 23S rRNA in the peptidyl transferase centre of the ribosome. *Nature* **377**, 309–314 (1995).
30. Kim, D. F. & Green, R. Base-pairing between 23S rRNA and tRNA in the ribosomal A site. *Mol. Cell* **4**, 859–864 (1999).
31. Muth, G. W., Ortoleva-Donnelly, L. & Strobel, S. A. A single adenosine with a neutral pKa in the ribosomal peptidyl transferase center. *Science* **289**, 947–950 (2000).
32. Polacek, N., Gaynor, M., Yassin, A. & Mankin, A. S. Ribosomal peptidyl transferase can withstand mutations at the putative catalytic nucleotide. *Nature* **411**, 498–501 (2001).
33. Thompson, J. et al. Analysis of mutations at residues A2451 and G2447 of 23S rRNA in the peptidyltransferase active site of the 50S ribosomal subunit. *Proc. Natl Acad. Sci. USA* **98**, 9002–9007 (2001).
34. Katunin, V. I., Muth, G. W., Strobel, S., Wintermeyer, W. & Rodnina, M. V. Important contribution to catalysis of peptide bond formation by a single ionizing group within the ribosome. *Mol. Cell* (in the press).
35. Bayfield, M. A., Dahlberg, A. E., Schulmeister, U., Dorner, S. & Barta, A. A conformational change in the ribosomal peptidyl transferase center upon active/inactive transition. *Proc. Natl Acad. Sci. USA* **98**, 10096–10101 (2001).
36. Muth, G. W., Chen, L., Kosek, A. & Strobel, S. pH-dependent conformational flexibility within the ribosomal peptidyl transferase center. *RNA* **7**, 1403–1415 (2001).
37. Xiong, L., Polacek, N., Sander, P., Boettger, E. G. & Mankin, A. S. pKa of adenine 2451 in the ribosomal peptidyl transferase center remains elusive. *RNA* **7**, 1365–1369 (2001).
38. Carter, A. P. et al. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature* **407**, 340–348 (2000).
39. Cate, J. H., Yusupov, M. M., Yusupova, G. Z., Earnest, T. N. & Noller, H. F. X-ray crystal structures of 70S ribosome functional complexes. *Science* **285**, 2095–2104 (1999).
40. Yusupov, M. M. et al. Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**, 883–896 (2001).
41. Ogle, J. M. et al. Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science* **292**, 897–902 (2001).
42. Yusupova, G. Z., Yusupov, M. M., Cate, J. H. D. & Noller, H. F. The path of messenger RNA through the ribosome. *Cell* **106**, 231–241 (2001).
43. Shine, J. & Dalgarno, L. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl Acad. Sci. USA* **71**, 1342–1346 (1974).
44. Steitz, J. A. & Jakes, K. How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *E. coli*. *Proc. Natl Acad. Sci. USA* **71**, 1342–1346 (1975).
45. Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B. & Steitz, T. A. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl Acad. Sci. USA* **98**, 4899–4903 (2001).

46. Stark, H. *et al.* Visualization of elongation factor Tu on the *Escherichia coli* ribosome. *Nature* **389**, 403–406 (1997).
47. Agrawal, R. K. *et al.* Visualization of tRNA movements on the *Escherichia coli* 70 S ribosome during the elongation cycle. *J. Cell Biol.* **150**, 447–459 (2000).
48. Wool, I. G., Correll, C. C. & Chan, Y.-L. in *The Ribosome. Structure, Function, Antibiotics, and Cellular Interactions* (eds Garrett, R. A. *et al.*) 461–473 (ASM Press, Washington, DC, 2000).
49. Ban, N. *et al.* Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit. *Nature* **400**, 841–847 (1999).
50. Agrawal, R. K., Linde, J., Sengupta, J., Nierhaus, K. H. & Frank, J. Localization of L11 protein on the ribosome and elucidation of its involvement in EF-G-dependent translocation. *J. Mol. Biol.* **311**, 777–787 (2001).
51. Czworkowski, J. & Moore, P. B. The elongation phase of protein synthesis. *Prog. Nucleic Acids Res. Mol. Biol.* **54**, 293–332 (1996).
52. Frank, J. & Agrawal, R. K. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature* **406**, 318–322 (2000).
53. Agrawal, R. K., Heagle, A. B., Penczek, P., Grassucci, R. A. & Frank, J. EF-G-dependent GTP hydrolysis induces translocation accompanied by large conformational changes in the 70S ribosome. *Nature Struct. Biol.* **6**, 643–647 (1999).
54. Stark, H., Rodnina, M. V., Wieden, H.-J., van Heel, M. & Wintermeyer, W. Large-scale movement of elongation factor G and extensive conformational change of the ribosome during translocation. *Cell* **100**, 301–309 (2000).
55. Peske, F., Matassova, N. B., Savelsbergh, A., Rodnina, M. V. & Wintermeyer, W. Conformationally restricted elongation factor G retains GTPase activity but is inactive in translocation on the ribosome. *Mol. Cell* **6**, 501–505 (2000).
56. Carter, A. P. *et al.* Crystal structure of an initiation factor bound to the 30S ribosomal subunit. *Science* **291**, 498–501 (2001).
57. Pioletti, M. *et al.* Crystal structures of complexes of the small ribosomal subunit with tetracycline, edeine and IF3. *EMBO J.* **20**, 1829–1839 (2001).
58. McCutcheon, J. P. *et al.* Location of translational initiation factor IF3 on the small ribosomal subunit. *Proc. Natl Acad. Sci. USA* **96**, 4301–4306 (1999).
59. Dallas, A. & Noller, H. F. Interaction of translation initiation factor 3 with the 30S ribosomal subunit. *Mol. Cell* **8**, 855–864 (2001).
60. Schluenzen, F. *et al.* Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. *Nature* **413**, 814–821 (2001).
61. Hansen, J. L., Ban, N., Nissen, P., Moore, P. B. & Steitz, T. A. The structures of four macrolide antibiotics bound to the large ribosomal subunit. *Mol. Cell* (in the press).
62. Beckmann, R. *et al.* Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* **107**, 361–372 (2001).
63. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).

Acknowledgements

We have benefited from discussions with our colleagues S. A. Strobel, J. L. Hansen, D. J. Klein and T. M. Schmeing, but the opinions expressed here are our responsibility. This work was supported by the Howard Hughes Medical Institute, NIH and the Agouron Institute.

Alternative pre-mRNA splicing and proteome expansion in metazoans

Tom Maniatis & Bosiljka Tasic

Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA

The protein coding sequences of most eukaryotic messenger RNA precursors (pre-mRNAs) are interrupted by non-coding sequences called introns. Pre-mRNA splicing is the process by which introns are removed and the protein coding elements assembled into mature mRNAs. Alternative pre-mRNA splicing selectively joins different protein coding elements to form mRNAs that encode proteins with distinct functions, and is therefore an important source of protein diversity. The elaboration of this mechanism may have had a significant role in the expansion of metazoan proteomes during evolution.

Metazoans display an extraordinarily broad spectrum of functional and behavioural complexity. Complex organisms probably evolved by increasing the number of components that constitute them (for example, proteins), and/or by elaborating the relationships between components (for example, regulatory networks). The size of the proteome of an organism (the complete set of proteins expressed by the genome during the life of an organism) can be expanded during evolution by increasing the number of genes, by elaborating pre-existing mechanisms that generate protein diversity, and by inventing new mechanisms. For example, the mechanism of somatic-cell DNA rearrangement, which can generate virtually unlimited diversity of antibodies and T-cell receptors, arose during vertebrate evolution.

Mechanisms that increase protein diversity in all metazoans include the use of multiple transcription start sites¹, alternative pre-mRNA splicing^{2–5}, polyadenylation⁶, pre-mRNA editing⁷, and post-translational protein modifications⁸. Among these mechanisms, alternative pre-mRNA splicing is considered to be the most important source of protein diversity in vertebrates^{2,3}. Here we review the mechanisms involved in the regulation of alternative pre-mRNA splicing, and discuss the effect of this process on expansion of the proteomes of multicellular organisms.

Exon recognition in constitutively spliced pre-mRNAs

The pre-mRNA splicing reaction is carried out by spliceosomes — multicomponent ribonucleoprotein complexes containing five small nuclear RNAs (snRNAs) and a large number of associated proteins^{9–11}. Spliceosomes recognize 5' and 3' splice sites, which are located at exon–intron boundaries (Fig. 1). The assembly of spliceosomes is a highly dynamic process, culminating in the juxtaposition of 5' and 3' splice sites in the catalytic core of the complex. The splicing reaction occurs via a two-step mechanism. In the first step, the 5' end of the intron is joined to an adenine residue in the branchpoint sequence upstream from the 3' splice site to form a branched intermediate called an intron lariat. In the second step, the exons are ligated and the intron lariat is released¹².

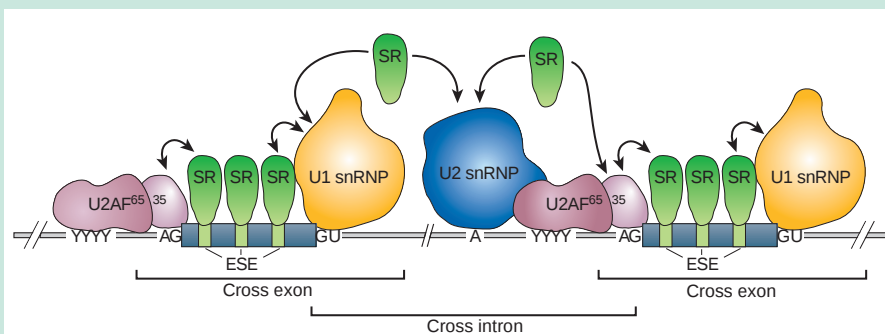
A fundamental problem in pre-mRNA splicing is 'exon recognition', the process by which exons are distinguished from introns, and intron–exon boundaries are accurately defined^{13,14}. The average size of a human exon is

150 nucleotides, whereas introns average around 3,500 nucleotides¹⁵, and can be as large as 500,000 nucleotides¹⁶. Thus, the splicing machinery must recognize small exon sequences located within vast stretches of intronic RNA. Moreover, 5' and 3' splice sites are poorly conserved, and introns contain large numbers of cryptic splice sites, which match the loose 5' or 3' splice-site consensus. Cryptic splice sites are normally avoided by the splicing machinery, but can be selected for splicing when normal splice sites are altered by mutation. Identification of the correct splice sites is achieved by virtue of their proximity to exons^{13,14}. Specific sequence elements in exons known as exonic splicing enhancers (ESEs) interact with SR proteins, a family of conserved serine/arginine-rich splicing factors^{14,17}. These recruit the splicing machinery to the flanking 5' and 3' splice sites (Fig. 1). Thus, exon sequences are under multiple evolutionary constraints, as they must be conserved not only for protein coding but also for recognition by SR proteins^{18,19}. The many cryptic splice sites present in introns could also be avoided by the presence of intronic splicing silencers (ISSs) and competition with splice sites flanking the recognized exons²⁰.

Once exon recognition is completed, the flanking splice sites must be joined in the correct 5'→3' order to prevent exon skipping. This is probably accomplished, at least in part, through the mechanistic coupling of transcription and splicing²¹. In this model, splicing factors, which are bound to the carboxy-terminal domain (CTD) of RNA polymerase II, interact with exons as they emerge from the exit pore of the polymerase. This interaction tethers the newly synthesized exon to the CTD until the next exon is synthesized. In large introns, many hours could pass between the synthesis of the 5' and 3' splice sites. Although coupling transcription to splicing can prevent exon skipping in constitutively spliced pre-mRNAs, exon skipping does occur during alternative pre-mRNA splicing. In this case, the presence or absence of regulatory proteins can determine whether an exon is recognized and subsequently included in the mature mRNA.

Mutations that interfere with proper exon recognition result in a large number of human genetic diseases²². Approximately 15% of the single base-pair mutations that cause human genetic diseases result in pre-mRNA splicing defects²³. Some of these mutations interfere with the function of normal 5' and 3' splice sites, thereby leading to the recognition of nearby pre-existing cryptic splice sites. Others actually create new splice sites that are used instead of

Figure 1 Exon recognition. The correct 5' (GU) and 3' (AG) splice sites are recognized by the splicing machinery on the basis of their proximity to exons. The exons contain exonic splicing enhancers (ESEs) that are binding sites for SR proteins. When bound to an ESE, the SR proteins recruit U1 snRNP to the downstream 5' splice site, and the splicing factor U2AF (65 and 35 kDa subunits) to the pyrimidine tract (YYYY) and the AG dinucleotide of the upstream 3' splice site, respectively. In turn, U2AF recruits U2 snRNP to the branchpoint sequence (A). Thus, the bound SR proteins recruit splicing factors to form a 'cross-exon' recognition complex. SR proteins also function in 'cross-intron' recognition by facilitating the interactions between U1 snRNP bound to the upstream 5' splice site and U2 snRNP bound to the branchpoint sequence.



the normal ones. Finally, single base-pair mutations within exons can interfere with the binding of SR proteins, leading to exon exclusion from the mature mRNA^{22,24,25}. For example, a translationally silent C-to-T mutation that occurs within an ESE of the human *survival of motor neuron 2 (SMN2)* gene disrupts the binding site of the SR protein SF2/ASF and leads to exon skipping²².

General mechanisms of alternative splicing

Alternative pre-mRNA splicing is the process by which multiple mRNAs can be generated from the same pre-mRNA by the differential joining of 5' and 3' splice sites. For example, exons can be extended or shortened, skipped or included, and introns can be removed or retained in the mRNA²². In some cases, exons are included in the mRNA in a mutually exclusive manner. Although there are very few examples in which the mechanisms of alternative splicing are known in detail, a general outline is understood. Regulatory proteins interact with specific sequences within pre-mRNAs and subsequently stimulate or repress exon recognition. These proteins bind directly to 5' or 3' splice sites, or to other pre-mRNA sequences called exonic or intronic splicing enhancers (ESEs or ISEs) and silencers (ESSs or ISSs). Enhancers and silencers stimulate or repress splice-site selection, respectively^{9,17,22,26–28}. A common feature of proteins that regulate splicing is the presence of two functional domains, an RNA-binding domain and a protein–protein interaction domain. The best characterized RNA-binding domains are the RNA-recognition motif (RRM) and K-homology (KH) domains²⁹. The three-dimensional structures of the two domains are distinct, as are the general features of the RNA-binding sites they recognize.

Most ESEs are recognized by SR proteins, which contain one or more RRM domains and an arginine/serine-rich (RS) protein–protein interaction domain^{27,30}. SR proteins are essential, multifunctional splicing factors required at different steps of spliceosome assembly¹⁷. They are also thought to mediate cross-intron interactions between splicing factors bound to the 5' and 3' splice sites¹⁷ (Fig. 1). Finally, SR proteins are required for cross-exon interactions in both constitutively and alternatively spliced pre-mRNAs^{14,17,27,30}.

The best characterized ESSs and ISSs are recognized by members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family. hnRNP proteins are highly abundant RNA-binding proteins that lack an RS domain. Several family members contain an arginine/glycine-rich domain that may be involved in both RNA binding and interactions with other proteins. Among the best characterized members of this family are the hnRNP A1 (ref. 31) and hnRNP I (ref. 32) proteins. hnRNP proteins are diffusely distributed throughout the nucleus^{33,34}, unlike SR proteins, which co-localize with other splicing factors in nuclear speckles^{27,30}.

It is important to note that some proteins containing an RS domain function as splicing repressors, whereas certain hnRNP proteins have been shown to function as splicing activators²⁸. SR proteins

were originally defined by a set of common functional characteristics¹⁷. Subsequently, other proteins containing RS domains that do not share these characteristics were discovered, and they are referred to as SR-like proteins. For example, two recently identified SR-like proteins have been shown to function as splicing repressors, and were therefore named SR-repressor proteins³⁵. We also note that there are splicing regulators that do not belong to either the SR or the hnRNP protein family. Some of these will be discussed below.

Remarkably, the role of regulatory proteins in splice-site selection can be affected by the promoter that generates the pre-mRNA³⁶. Thus transcription of the same pre-mRNA from different promoters can produce distinct mRNAs. This mechanism of alternative splicing could be a consequence of the coupling between transcription and pre-mRNA splicing²¹. For example, particular SR proteins could be differentially recruited to RNA polymerase complexes assembled on different promoters, and then transferred to cognate splicing enhancers to promote the inclusion of specific exons⁵.

Well characterized examples of alternative splicing

The mechanism of exon recognition in constitutively spliced pre-mRNAs provides the basis for positive and negative regulation of alternative splicing. The organization of regulatory sequences within pre-mRNAs (ESEs, ESSs, ISEs and ISSs) and the relative ratios of different regulatory proteins determine which splice sites are used in the splicing reaction. This, in turn, determines which exons are included in the mRNA^{4,5,28}.

The best characterized examples of regulated alternative splicing derive from studies of the *Drosophila* sex-determination pathway³⁷. The key regulatory factors were first identified by genetic analyses^{38,39}, and then characterized in biochemical experiments. Multiple steps in this pathway are regulated by different mechanisms of alternative pre-mRNA splicing. The examples described below include both splice-site repression and activation.

The *Drosophila* female-specific protein Sex-lethal (SXL) represses male-specific 3' splice sites in *transformer (tra)* and *sxl* pre-mRNAs by two distinct mechanisms. As shown in Fig. 2a, exon 2 of *tra* pre-mRNA is preceded by two alternative 3' splice sites. The proximal and distal 3' splice sites are used in males and females, respectively. Splicing to the male-specific 3' splice site produces an mRNA containing a premature translational stop codon. By contrast, splicing to the female-specific 3' splice site produces an mRNA that encodes functional TRA protein. In males, the splicing factor U2AF binds to the male-specific 3' splice site and initiates spliceosome assembly (Fig. 2a)³⁷. However in females, this splice site is bound by the female-specific splicing repressor SXL, thus blocking the binding of U2AF. Instead, U2AF binds to the female-specific 3' splice site, and functional *tra* mRNA is produced (Fig. 2a).

SXL also regulates the alternative splicing of its own pre-mRNA, albeit by an entirely different mechanism⁴⁰. *sxl* exon 3 is excluded

from the mRNA only in females. In males, inclusion of exon 3 introduces a premature translational stop codon. Exon 3 inclusion requires the protein SPF45, a second-step splicing factor that binds to the AG dinucleotide of the male-specific 3' splice site (Fig. 2b). In females, SXL binds to a site adjacent to SPF45, and the two proteins interact. This interaction interferes with the activity of SPF45, and thus blocks the second step of the splicing reaction. As a consequence, exon 3 is skipped and exon 2 is spliced to exon 4, thus producing an mRNA encoding functional SXL protein. Thus, SXL blocks the first step of the splicing reaction in *tra* pre-mRNA and the second step in *sxl* pre-mRNA. *sxl* autoregulation is the only known example in which alternative pre-mRNA splicing is regulated at the second step of the splicing reaction.

While the previous examples of regulated alternative splicing involve splice-site repression, the female-specific splicing of *Drosophila doublesex* (*dsx*) pre-mRNA is the best characterized example of splice-site activation³⁷. The 3' splice site immediately upstream from exon 4 of *dsx* pre-mRNA is not recognized by the splicing machinery in males, thus leading to the exclusion of this exon (Fig. 2c). The male-specific *dsx* mRNA encodes a transcriptional repressor of female-specific genes. In females, the regulatory protein TRA promotes the cooperative binding of an SR protein, RBP1, and an SR-like protein, Transformer 2 (TRA2), to individual ESEs within exon 4 (refs 41, 42). This heterotrimeric protein complex recruits the splicing machinery to the upstream 3' splice site, leading to the inclusion of exon 4. The female-specific *dsx* mRNA encodes a transcriptional repressor of male-specific genes.

The basic mechanisms of alternative splicing established in *Drosophila* have been shown to function in mammals. Alternative splice-site selection in mammals is also controlled by differential binding of regulatory proteins to splice sites, or to enhancer or

silencer sequences within the pre-mRNA²⁸. The organization of these sequences and the interplay of different regulatory proteins determine the outcome of the splicing reaction. As in *Drosophila*, regulatory proteins can exert competing influences on the splicing of a pre-mRNA.

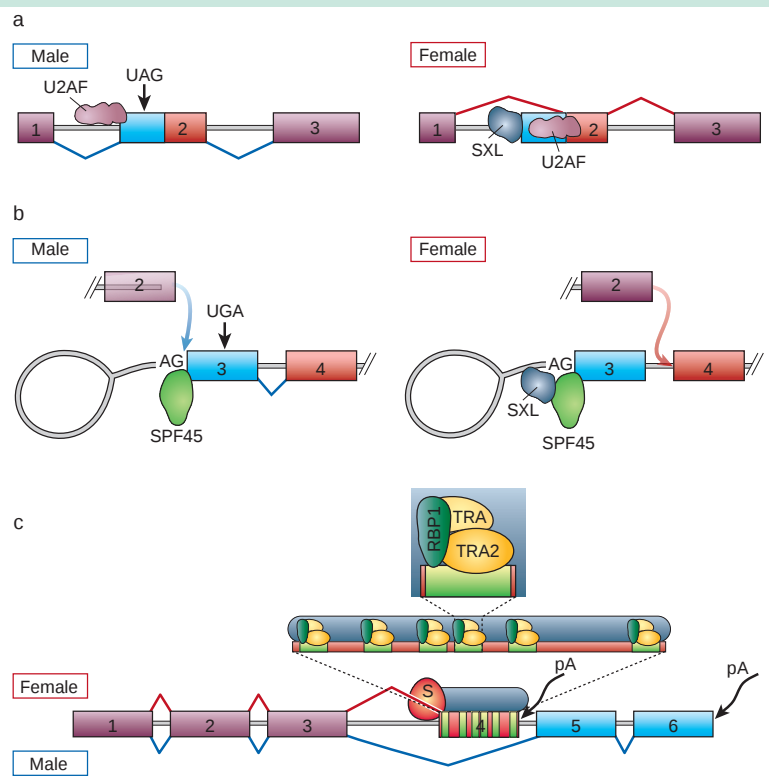
For example, the binding of hnRNP A1 to an ESS in an exon of HIV *tat* pre-mRNA represses the inclusion of this exon in the mRNA. The binding of hnRNP A1 prevents the interaction of the SR protein SC35 with a nearby ESE³¹. Although the ESS and the ESE do not directly overlap, hnRNP A1 bound to the ESS seems to promote cooperative binding of additional hnRNP A1 proteins to adjacent exon sequences, thus spreading into the ESE. However, another SR protein, SF2/ASF, has a higher affinity for the ESE than SC35 and is therefore able to displace hnRNP A1.

One of the best characterized mammalian repressors of exon recognition is hnRNP I, also called a polypyrimidine tract-binding protein (PTB)^{22,27,28,43}. This protein can repress exon inclusion by directly interfering with binding of general splicing factors to the pyrimidine tract. However, in most cases, the hnRNP I/PTB-binding sites flank the excluded exon. Two mechanisms for repression by hnRNP I/PTB have been proposed to account for this observation³². In the first mechanism, hnRNP I/PTB proteins bound on both sides of the exon interact with each other in such a way that the intervening exon 'loops out' and is isolated from the splicing machinery. In the second mechanism, hnRNP I/PTB cooperatively spreads across the exon, thus creating a 'zone of silencing'^{22,32}.

Regulation of tissue-specific alternative splicing

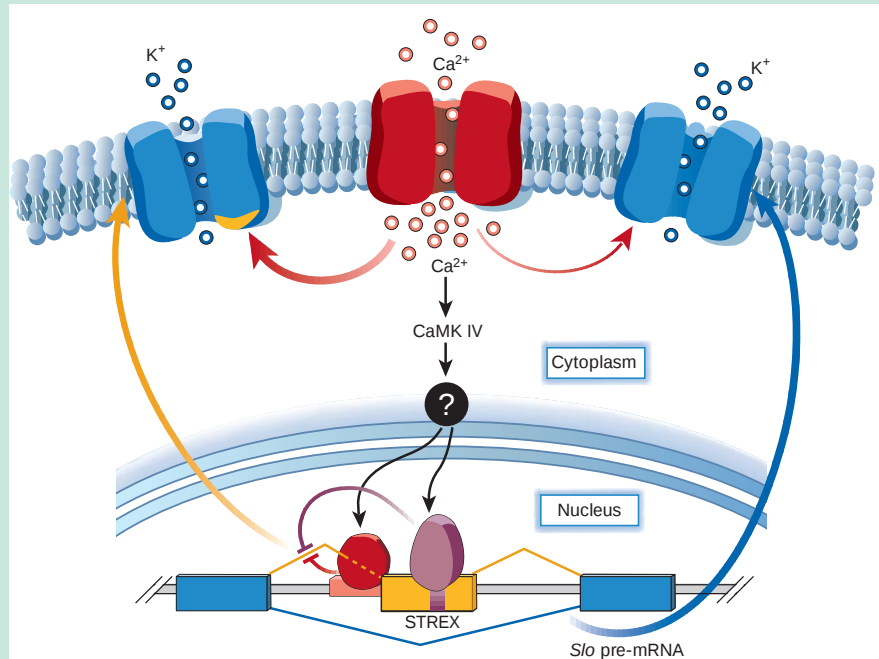
In addition to the ubiquitous RNA-binding proteins implicated generally in alternative splicing, a number of specific regulatory proteins have been identified, and some of them are expressed only

Figure 2 Regulation of alternative pre-mRNA splicing in the *Drosophila* sex-determination pathway. **a**, Alternative selection of 3' splice sites preceding exon 2 of *tra* pre-mRNA is regulated by the SXL protein. In males, the splicing factor U2AF binds to the proximal 3' splice site, leading to an mRNA containing a premature translational stop codon (UAG). In females, SXL binds to the proximal 3' splice site, thus preventing the binding of U2AF. Instead, U2AF binds to the distal 3' splice site, leading to an mRNA that encodes functional TRA protein. In all panels, the exons are indicated by coloured rectangles, while introns are shown as pale grey lines. **b**, Alternative inclusion of exon 3 of *sxl* pre-mRNA is regulated by SXL protein. In both males and females, the first step of the splicing reaction results in lariat formation at the branchpoint sequence upstream from the 3' splice site preceding exon 3. Subsequently, the second-step splicing factor SPF45 binds to the AG dinucleotide of this splice site. In males, SPF45 promotes the second step of the splicing reaction, leading to the inclusion of exon 3. In females, SXL binds to a sequence upstream of the AG dinucleotide, interacts with SPF45 and inhibits its activity. This prevents the second step of the splicing reaction, leading to the exclusion of exon 3 and splicing of exon 2 to exon 4. Seven constitutively spliced exons are not shown. **c**, Alternative splicing of *dsx* pre-mRNA is regulated by the assembly of heterotrimeric protein complexes on female-specific ESEs. The first three exons are constitutively spliced in both sexes. In males, the 3' splice site preceding exon 4 is not recognized by the splicing machinery, resulting in the exclusion of this exon, and splicing of exon 3 to exon 5. In females, the female-specific TRA protein promotes the binding of the SR protein RBP1, and the SR-like protein TRA2 to six copies of an ESE (indicated by green rectangles). These splicing enhancer complexes then recruit the splicing machinery to the 3' splice site



preceding exon 4, leading to its inclusion in the mRNA. In females, polyadenylation (pA) occurs downstream of exon 4, whereas in males it occurs downstream of exon 6. 'S' designates the splicing machinery.

Figure 3 Inducible alternative splicing of rat *Slo* pre-mRNA. The fraction of *Slo* mRNAs containing an exon called STREX (stress axis-regulated exon, shown in yellow) is regulated by neuronal activity. Depolarization of the plasma membrane increases the intracellular Ca^{2+} concentration, leading to the activation of CaMK IV. This, in turn, is thought to result in the binding of repressor proteins to silencer sequences, one located upstream and the other within STREX. These silencers are shown as red rectangles with hypothetical repressors bound to them. In the presence of the repressors, the STREX exon is excluded from *Slo* mRNA. The alternatively spliced mRNAs encode two protein isoforms (shown in blue). The channel encoded by mRNA containing the STREX exon has an additional (yellow) domain, and is more sensitive to intracellular Ca^{2+} concentration. For clarity, only the portion of *Slo* pre-mRNA containing the STREX exon is shown.



in certain tissues. These proteins regulate alternative splicing of specific sets of pre-mRNAs.

Drosophila ELAV (for embryonic-lethal, abnormal visual system) is a pan-neuronal pre-mRNA-binding protein that regulates alternative splicing of at least three different pre-mRNAs⁴⁴. This protein contains three RRM, which are required for binding to uridine-rich regulatory sequences in the target pre-mRNAs. For example, ELAV binding to *neuroglian* pre-mRNA results in alternative inclusion of a neural-specific terminal exon. As yet there is no evidence that the mammalian homologues of *Drosophila* ELAV are involved in the regulation of pre-mRNA splicing^{45,46}.

The *Drosophila* Half pint (HFP) protein regulates alternative splicing of a subset of pre-mRNAs in the ovary⁴⁷. One of these encodes the ovarian tumour protein, which is required for oogenesis, while another encodes a translation initiation factor. HFP is also required for constitutive splicing of the *gurken* pre-mRNA in oocytes. The mammalian orthologues of HFP are the human PUF60 protein⁴⁸ and the rat Siah-binding protein⁴⁹. These proteins interact with the splicing factor U2AF⁶⁵, which binds to 3' splice sites in all pre-mRNAs (Fig. 1). Although HFP targets specific pre-mRNAs, human PUF60 stimulates splicing nonspecifically in mammalian cell extracts, suggesting that it is a general splicing factor⁴⁸.

The *Drosophila* P-element transposon is active only in the germline because the splicing of its transposase pre-mRNA is repressed in somatic cells. The P-element somatic inhibitor protein (PSI) binds to the transposase pre-mRNA and inhibits removal of a specific intron in somatic cells, leading to the production of a truncated protein that represses transposition⁵⁰. By contrast, PSI is not expressed in germ cells, and functional transposase is produced. PSI binds to 'pseudo' 5' splice sites in transposase pre-mRNA and recruits U1 snRNP to these sites through interactions with the U1 snRNP 70K protein⁵⁰. This leads to the formation of an abortive 5' splice-site complex that interferes with splicing at the normal 5' splice site. The mammalian homologue of PSI, called KSRP (for KH-type splicing regulatory protein), regulates neural-specific splicing of the human *src* pre-mRNA⁵¹.

The mammalian splicing factor NOVA-1 (for neuro-oncological ventral antigen-1) regulates alternative pre-mRNA splicing in neurons. NOVA-1 binds to specific intronic sequences of target pre-mRNAs, and stimulates the inclusion of specific exons in the

corresponding mRNAs⁴⁵. For example, inclusion of alternatively spliced exons in the pre-mRNAs encoding subunits of the glycine and GABA (γ -aminobutyric acid) receptors requires NOVA-1.

At present, there are only a few examples of proteins that regulate tissue-specific alternative splicing, but it is likely that many more regulatory proteins, target pre-mRNAs and regulatory mechanisms will be discovered. Although many tissue-specific regulatory proteins are conserved between flies and mammals, their target pre-mRNAs can be different. This suggests a role for tissue-specific alternative splicing in the generation of species-specific traits and functions.

Inducible alternative splicing

The proteome of a cell can rapidly change in response to extracellular stimuli through complex signal-transduction pathways. Changes in protein composition can be regulated at many different levels, but have been studied primarily at the level of transcription and post-translational protein modification. Recently, several cases of inducible alternative pre-mRNA splicing have been reported.

An example from the mammalian immune system is the alternative splicing of a pre-mRNA encoding distinct isoforms of CD45, a transmembrane protein tyrosine phosphatase. CD45 pre-mRNA is alternatively spliced in response to T-cell activation, leading to the production of proteins with distinct extracellular domains. Recent studies revealed a role for protein kinase C and Ras in the signalling pathway required for the switch in CD45 alternative splicing⁵². Although the details of the signalling pathway are not fully understood, several lines of evidence point to SR proteins as the downstream effectors in the pathway. First, the relative levels of several SR proteins change upon T-cell activation⁵³. Second, overexpression of different SR proteins in cultured cells leads to distinct patterns of alternative splicing of CD45 pre-mRNA⁵³. Finally, a T-cell-specific conditional knockout of the gene encoding the SR protein SC35 interferes with the normal pattern of CD45 alternative splicing upon T-cell activation⁵⁴. Based on these observations, it has been proposed that T-cell activation leads to the production of a new combination of SR proteins that bind to CD45 pre-mRNA and change the pattern of its alternative splicing⁵²⁻⁵⁴.

Similar mechanisms may control alternative splicing in the brain in response to neuronal activity⁴³. For example, alternative splicing of

the rat SR-like protein Tra2- β (an orthologue of *Drosophila* TRA2) changes in response to small molecule-induced neural activity⁵⁵. These changes in the types and levels of Tra2- β isoforms could, in turn, control the alternative splicing of other brain pre-mRNAs. A potential target of human Tra2- β is the *SMN2* pre-mRNA, because its splicing pattern can be altered by overexpression of this SR-like protein⁵⁶.

The *ania-6* pre-mRNA, which encodes a member of a new family of cyclins, provides another example of neuronal activity-dependent alternative pre-mRNA splicing. The synthesis of this protein is inducible in the adult brain of rats by cocaine and dopamine agonists, and two distinct *ania-6* mRNAs are generated by alternative splicing in response to different neurotransmitters and drugs⁵⁷. One mRNA encodes a protein that associates with RNA polymerase II and co-localizes with splicing factors in nuclear speckles, while the other mRNA encodes a protein that localizes to the cytoplasm⁵⁷. Thus, two functionally distinct proteins are produced from the same pre-mRNA in response to neural activity.

The only case of inducible alternative splicing in which the regulatory sequences within pre-mRNA have been identified is the neural activity-dependent alternative splicing of the rat *Slo* pre-mRNA⁵⁸. This pre-mRNA is alternatively spliced, leading to the production of multiple protein isoforms of a calcium-dependent potassium channel^{3,59}. These isoforms display distinct electrophysiological properties. A cell-culture model system was used to study the signalling pathway that induces *Slo* alternative splicing, and to identify *cis*-acting regulatory sequences required for the inclusion of a particular exon, stress axis-regulated exon (STREX), in *Slo* mRNA⁵⁸. Depolarization of pituitary cells represses inclusion of the STREX exon in a process that requires the activity of a Ca²⁺/calmodulin-dependent protein kinase (CaMK IV). This signalling pathway acts through an ESS in the STREX exon and the 3' splice site upstream of STREX. Thus, it seems that depolarization of the membrane leads to the activation of CaMK IV, which in turn activates unidentified repressor proteins that bind the STREX exon and the 3' splice site, thereby preventing STREX exon inclusion in *Slo* mRNA⁵⁸ (Fig. 3). Inclusion of the STREX exon increases the calcium sensitivity of the channel, which can modulate the electrical properties of the cell⁵⁸.

Alternative trans-splicing

In the examples of alternative splicing discussed above, exons located within an individual pre-mRNA are differentially joined to generate mature mRNAs (alternative *cis*-splicing). *Trans*-splicing can join exons located within separate pre-mRNAs, whereas alternative *trans*-splicing differentially uses exons within separate pre-mRNAs to produce distinct mRNAs. *Trans*-splicing was first discovered in trypanosomes and later found in organisms as complex as chordates⁶⁰. However, this type of *trans*-splicing, called spliced leader (SL)-addition *trans*-splicing, involves specialized SL RNAs that provide 5'-terminal non-coding exons for all mRNAs⁶¹. SL-addition *trans*-splicing is not a source of protein diversity, as it is a constitutive process that does not lead to the production of alternatively spliced mRNAs. So far, there is no evidence for SL-addition *trans*-splicing in vertebrates.

The potential for mammalian *trans*-splicing was first demonstrated *in vitro*¹⁴, and then shown to occur *in vivo* between separate viral pre-mRNAs, and between viral and cellular RNAs in infected cells^{62,63}. Alternative *trans*-splicing has also been shown to result in exon duplications in several mammalian cellular RNAs^{64–67}. However, mRNAs containing duplicated exons have not been shown to encode functional proteins, and in one case, exon duplication is not conserved in closely related organisms⁶⁸.

Although several examples of mammalian interchromosomal *trans*-splicing have been reported, none of these studies definitively proves the existence of this phenomenon^{69–72}. There are also cases of intergenic *trans*-splicing between pre-mRNAs encoded by closely linked genes in mammals^{73–76}. For example, several hybrid mRNAs are produced from a cluster of human cytochrome *P450* 3A genes⁷⁶

(Fig. 4a). Three of the *P450* genes (here labelled as 2, 3 and 4) are transcribed from one DNA strand, whereas one (1) is transcribed from the opposite DNA strand. Hybrid mRNAs containing exon 1 of gene 1 joined to various exons of genes 2 and 4 were detected (Fig. 4a). Because gene 1 is transcribed from one DNA strand and genes 2 and 4 from the other, the hybrid mRNAs must be generated by *trans*-splicing. However, the hybrid intergenic mRNAs are produced at levels that are orders of magnitude lower than those of intragenic mRNAs, and the existence of the endogenous proteins encoded by these hybrid mRNAs has not been demonstrated⁷⁶.

It is important to note that in every example of mammalian *trans*-splicing so far reported the pre-mRNAs also engage in *cis*-splicing. Therefore, it is possible that the observed *trans*-splicing represents 'splicing noise' resulting from low levels of splice-site pairing between separate pre-mRNAs. Thus, the functional significance of mammalian alternative *trans*-splicing remains to be established.

The only known example in which alternative *trans*-splicing is required for the production of an essential protein occurs in the *Drosophila* modifier of *mdg4* (*mod(mdg4)*) locus^{77,78}. MOD(MDG4) protein isoforms, which function in the establishment or maintenance of chromatin structure, are encoded by an unusually complex genetic locus (Fig. 4b). Twenty-six alternatively spliced mRNAs encoded by this locus have been identified, and they all contain four common exons located at the 5' end of the locus⁷⁸ (shown in red in Fig. 4b). Distinct mRNAs are generated by alternative splicing of the fourth common exon to individual downstream 'variable' exons (shown in blue or yellow in Fig. 4b). Strikingly, a number of variable exons are transcribed from the opposite DNA strand (blue exons in Fig. 4b), strongly suggesting that they are joined to the fourth common exon by *trans*-splicing.

A central question in the regulation of alternative *trans*-splicing is how splice sites located on separate pre-mRNAs are recognized by the splicing machinery and correctly joined. This is not an entirely new question, as splice sites at the ends of large introns are essentially configured *in trans*. In this case, the correct joining of splice sites is probably achieved by coupling transcription and splicing through interactions between the splicing machinery and the CTD of RNA polymerase II (ref. 21). In this mechanism, the splicing factors connect the 5' splice site and the CTD as the long intron is being transcribed. The proximity of the eventually synthesized 3' splice site to the CTD would ensure an interaction between splicing complexes assembled on the 5' and 3' splice sites.

This type of coupling cannot occur in *trans*-splicing because different RNA polymerase complexes transcribe separate *trans*-splicing precursors. Therefore, there must be another mechanism for bringing together the 5' and 3' splice sites. This might be achieved by the localized transcription of pre-mRNAs in the nucleus — by restricting transcription to 'gene expression factories'²¹ only those pre-mRNAs transcribed in the same factory would engage in *trans*-splicing. This could explain the close linkage of *trans*-spliced exons in the *mod(mdg4)* locus, and the previously mentioned cases of intergenic *trans*-splicing. Nuclear compartmentalization could also prevent inappropriate splicing to pre-mRNAs encoded by other genes. Finally, it is possible that *trans*-splicing precursors interact through specific base pairing, or through interactions between proteins bound to each of the precursors^{77,78}.

Comparative genomics and alternative pre-mRNA splicing

Alternative pre-mRNA splicing is an important source of protein diversity that may have contributed to the increase in the phenotypic complexity of metazoans during evolution^{2,3}. This proposal was based, in part, on the unexpectedly small difference in gene number in different organisms from yeast to humans. For example, flies have only about twice as many genes as yeast⁷⁹. In addition, worms have more genes than flies (19,000 and 14,000, respectively), but flies are clearly more complex in their development, morphology and behaviour. Remarkably, humans have only about 35,000 genes^{80,81}.

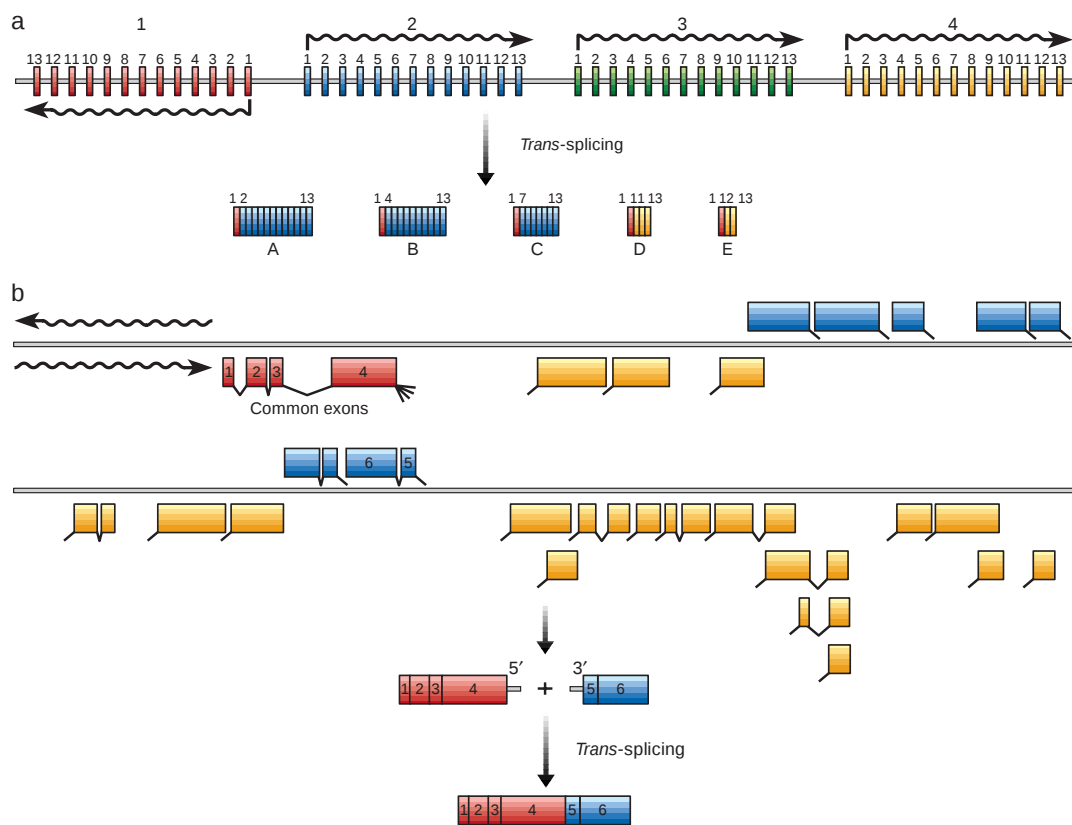


Figure 4 Alternative *trans*-splicing in mammals and flies. **a**, Four cytochrome P450 3A genes are arranged in a cluster that spans ~200 kilobases of genomic DNA. For simplicity, the genes have been designated 1–4 (they were designated *CYP3A43*, *CYP3A4*, *CYP3A7*, *CYP3A5*, respectively, in the original reference)⁷⁸. Gene 1 is transcribed from one DNA strand, whereas genes 2–4 are transcribed from the opposite strand. The direction of transcription on each strand is indicated by the wavy arrows. The small coloured boxes are exons, and the pale grey lines are introns. Hybrid intergenic mRNAs are produced by *trans*-splicing between exon 1 of gene 1 and various exons in genes 2 and 4 to generate the mRNAs labelled A–E. **b**, The *Drosophila mod(mdg4)* locus spans 28 kilobases, and encodes 26 alternatively

spliced mRNAs. Each mRNA is generated by splicing of four common exons (1–4, shown in red) to one of the exons located downstream (shown in yellow or blue). The yellow exons are transcribed from the same DNA strand as the common exons, while the blue exons are transcribed from the opposite DNA strand. The 5' splice site downstream from common exon 4 is indicated by the four slanted lines. Similarly, the 3' splice site upstream of each alternatively spliced exon is indicated by a single slanted line. An example of *trans*-splicing within this locus is shown at the bottom of the figure. Two *trans*-splicing precursors, containing exons 1–4 or 5 and 6, are transcribed from opposite DNA strands. The exons 4 and 5 are then joined by *trans*-splicing.

Although this number is still being debated⁸², even the highest estimate for humans is only around threefold greater than that for worms. It is unlikely that this difference alone could explain the obvious differences in functional and behavioural complexity between invertebrates and vertebrates. The increased organismic complexity of vertebrates could be a consequence of the elaboration of mechanisms that increase proteome size and the evolution of more extensive networks of gene regulation⁸³.

A measure of the relative contribution of alternative splicing to proteome size could be obtained by dividing the total number of distinct full-length complementary DNAs (cDNAs) resulting from alternative splicing by the number of genes of an organism. Unfortunately, relatively few full-length cDNA sequences are available^{84–86}, and the annotation of several genomes, including the human genome, is still imprecise. In the absence of this information, partial cDNA sequences and expressed sequence tags (ESTs) have been used to detect alternative splicing events^{87,88}. Efforts are also being made to use DNA microarrays to identify alternatively spliced mRNAs on a large scale^{89–91}. These approaches have provided preliminary estimates of the percentage of genes with alternatively spliced forms (%G_{ASF}) and the average number of alternatively spliced forms per

gene (ASF/G). The product of these two numbers provides a quantitative estimate of the contribution of alternative splicing to proteome size.

Estimates of %G_{ASF} in humans range from 35 to 59% (refs 87, 88). However, it is generally agreed that this is an underestimate, because %G_{ASF} depends on the number of ESTs examined⁹². It is important to note that most of the detected alternative splicing events (about 80%) lead to changes in the amino acid sequence of the encoded proteins⁸⁸. In a recent report, random sets of 650 non-redundant cDNAs were compared with 100,000 ESTs in different organisms ranging from worms and flies to humans⁹². The surprising outcome was that the estimate for %G_{ASF} is approximately the same in all organisms tested⁹².

Similar values have also been obtained for ASF/G in different organisms; however, the estimates were adversely affected if the analysis included 'outlier' genes that generate extraordinarily large numbers of alternatively spliced forms (ref. 92, and J. Valcarcel and P. Bork, personal communication). For example, the *Drosophila* axon guidance receptor gene, *Dscam*, contains 95 alternatively spliced exons organized into four clusters, and it has the potential to generate over 38,000 different protein isoforms — nearly three times more proteins than the number of genes in *Drosophila*⁹³. Remarkable

but less extensive examples in mammals are the previously mentioned *Slo* gene, and three *neurexin* genes, which encode proteins that may act as synaptic-cell-surface receptors or cell-adhesion molecules⁹⁴. The *Slo* gene and the three *neurexin* genes have the potential to generate over 500 and 2,000 alternatively spliced forms, respectively^{3,16}.

The conclusion of this analysis is that the relative contribution of alternative splicing to the size of the proteome does not seem to be different in evolutionarily divergent metazoans if 'outlier' genes are not included in the analysis⁹². However, it is possible that the differences will be observed once the 'outlier' genes are considered. To account for their contribution to the overall estimate of ASF/G, the organization of all genes and the sequence of all full-length cDNAs are required. Thus, the relationship between alternative splicing and proteome size of different metazoans remains to be established.

Perspectives

At least two forces affect the evolution of complex organisms. On the one hand there is an extraordinary pressure to conserve the basic components of cellular machines among organisms that differ widely in complexity^{80,81}. For example, the cellular machines required for different steps of gene expression, such as transcription⁹⁵, splicing¹¹ and mRNA export⁹⁶, are so highly conserved between yeast and humans that many components are interchangeable.

On the other hand, both the proteome size and the complexity of regulatory networks increase as more complex organisms evolve. Proteome size is determined by the number of genes in an organism, and by mechanisms that expand the coding capacity of the genome, including alternative pre-mRNA splicing. However, as discussed above, the relative contribution of alternative splicing to proteome expansion is not well understood. The organization of regulatory networks and their effect on organismic complexity are even more difficult to determine. Moreover, it is possible that a relatively small increase in the size of the proteome or a change in a key component of the regulatory networks could lead to dramatic changes in organismic complexity. If this were the case, the proteome size would not accurately correlate with organismic complexity. Rather, insights into the origins of organismic complexity would require an understanding of regulatory networks⁸³. □

- Quelle, D. E., Zindy, F., Ashmun, R. A. & Sherr, C. J. Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell* **83**, 993–1000 (1995).
- Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103**, 367–370 (2000).
- Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–107 (2001).
- Goldstrohm, A. C., Greenleaf, A. L. & Garcia-Blanco, M. A. Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene* **277**, 31–47 (2001).
- Caceres, J. F. & Kornblitt, A. R. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* **18**, 186–193 (2002).
- Gautheret, D., Poirat, O., Lopez, F., Audic, S. & Claverie, J. M. Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.* **8**, 524–530 (1998).
- Keegan, L. P., Gallo, A. & O'Connell, M. A. The many roles of an RNA editor. *Nature Rev. Genet.* **2**, 869–878 (2001).
- Banks, R. E. *et al.* Proteomics: new perspectives, new biomedical opportunities. *Lancet* **356**, 1749–1756 (2000).
- Reed, R. Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell Biol.* **12**, 340–345 (2000).
- Stevens, S. W. *et al.* Composition and functional characterization of the yeast spliceosomal pentas-nRNP. *Mol. Cell* **9**, 31–44 (2002).
- Zhou, Z., Licklider, L., Gygi, S. & Reed, R. The proteome of functional human spliceosomes. *Nature* (submitted).
- Staley, J. P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315–326 (1998).
- Berger, S. M. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**, 2411–2414 (1995).
- Reed, R. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr. Opin. Genet. Dev.* **6**, 215–220 (1996).
- Deutsch, M. & Long, M. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27**, 3219–3228 (1999).
- Rowen, L. *et al.* Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity. *Genomics* **79**, 587–597 (2002).
- Graveley, B. R. Sorting out the complexity of SR protein functions. *RNA* **6**, 1197–1211 (2000).
- Schaal, T. D. & Maniatis, T. Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol. Cell. Biol.* **19**, 261–273 (1999).
- Mayeda, A., Sreanot, G. R., Chandler, S. D., Fu, X. D. & Krainer, A. R. Substrate specificities of SR

- proteins in constitutive splicing are determined by their RNA recognition motifs and composite pre-mRNA exonic elements. *Mol. Cell. Biol.* **19**, 1853–1863 (1999).
- Sun, H. & Chasin, L. A. Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* **20**, 6414–6425 (2000).
- Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416**, 499–506 (2002).
- Cartegni, L., Chew, S. L. & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.* **3**, 285–298 (2002).
- Krawczak, M., Reiss, J. & Cooper, D. N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* **90**, 41–54 (1992).
- Valentine, C. R. The association of nonsense codons with exon skipping. *Mutat. Res.* **411**, 87–117 (1998).
- Liu, H. X., Cartegni, L., Zhang, M. Q. & Krainer, A. R. A mechanism for exon skipping caused by nonsense or missense mutations in *BRCA1* and other genes. *Nature Genet.* **27**, 55–58 (2001).
- Blencowe, B. J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25**, 106–110 (2000).
- Hastings, M. L. & Krainer, A. R. Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol.* **13**, 302–309 (2001).
- Smith, C. W. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* **25**, 381–388 (2000).
- Perez-Canadillas, J. M. & Varani, G. Recent advances in RNA-protein recognition. *Curr. Opin. Struct. Biol.* **11**, 53–58 (2001).
- Graveley, B. R., Hertel, K. J. & Maniatis, T. SR proteins are 'locators' of the RNA splicing machinery. *Curr. Biol.* **9**, R6–R7 (1999).
- Zhu, J., Mayeda, A. & Krainer, A. R. Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol. Cell* **8**, 1351–1361 (2001).
- Wagner, E. J. & Garcia-Blanco, M. A. Polypyrimidine tract binding protein antagonizes exon definition. *Mol. Cell. Biol.* **21**, 3281–3288 (2001).
- Dreyfuss, G., Matunis, M. J., Pinol-Roma, S. & Burd, C. G. hnRNP proteins and the biogenesis of mRNA. *Annu. Rev. Biochem.* **62**, 289–321 (1993).
- Krecic, A. M. & Swanson, M. S. hnRNP complexes: composition, structure, and function. *Curr. Opin. Cell Biol.* **11**, 363–371 (1999).
- Cowper, A. E., Caceres, J. F., Mayeda, A. & Sreanot, G. R. Serine-arginine (SR) protein-like factors that antagonize authentic SR proteins and regulate alternative splicing. *J. Biol. Chem.* **276**, 48908–48914 (2001).
- Cramer, P. *et al.* Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol. Cell* **4**, 251–258 (1999).
- Schutt, C. & Nothiger, R. Structure, function and evolution of sex-determining systems in Dipteran insects. *Development* **127**, 667–677 (2000).
- Baker, B. S. Sex in flies: the splice of life. *Nature* **340**, 521–524 (1989).
- Cline, T. W. & Meyer, B. J. Vive la différence: males vs females in flies vs worms. *Annu. Rev. Genet.* **30**, 637–702 (1996).
- Lallena, M. J., Chalmers, K. J., Llamazares, S., Lamond, A. I. & Valcarcel, J. Splicing regulation at the second catalytic step by Sex-lethal involves 3' splice site recognition by SPF45. *Cell* **109**, 285–296 (2002).
- Lynch, K. W. & Maniatis, T. Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila* doublesex splicing enhancer. *Genes Dev.* **10**, 2089–2101 (1996).
- Hertel, K. J. & Maniatis, T. The function of multisite splicing enhancers. *Mol. Cell* **1**, 449–455 (1998).
- Grabowski, P. J. & Black, D. L. Alternative RNA splicing in the nervous system. *Prog. Neurobiol.* **65**, 289–308 (2001).
- Lisbin, M. J., Qiu, J. & White, K. The neuron-specific RNA-binding protein ELAV regulates neuroglian alternative splicing in neurons and binds directly to its pre-mRNA. *Genes Dev.* **15**, 2546–2561 (2001).
- Dredge, B. K., Polydorides, A. D. & Darnell, R. B. The splice of life: alternative splicing and neurological disease. *Nature Rev. Neurosci.* **2**, 43–50 (2001).
- Toba, G., Qui, J., Koushika, S. P. & White, K. Ectopic expression of *Drosophila* ELAV and human HuD in *Drosophila* wing disc cells reveals functional distinctions and similarities. *J. Cell Sci.* **115**, 2413–2421 (2002).
- Van Buskirk, C. & Schupbach, T. *half pint* regulates alternative splice site selection in *Drosophila*. *Dev. Cell* **2**, 343–353 (2002).
- Page-McCaw, P. S., Amonlirdviman, K. & Sharp, P. A. PUF60: a novel U2AF65-related splicing activity. *RNA* **5**, 1548–1560 (1999).
- Polev, A., Hartmann, A. & Stamm, S. A trans-acting factor, isolated by the three-hybrid system, that influences alternative splicing of the amyloid precursor protein minigene. *Eur. J. Biochem.* **267**, 4002–4010 (2000).
- Labourier, E., Adams, M. D. & Rio, D. C. Modulation of P-element pre-mRNA splicing by a direct interaction between PSI and U1 snRNP 70K protein. *Mol. Cell* **8**, 363–373 (2001).
- Min, H., Turck, C. W., Nikolic, J. M. & Black, D. L. A new regulatory protein, KSRP, mediates exon inclusion through an intronic splicing enhancer. *Genes Dev.* **11**, 1023–1036 (1997).
- Lynch, K. W. & Weiss, A. A model system for activation-induced alternative splicing of CD45 pre-mRNA in T cells implicates protein kinase C and Ras. *Mol. Cell. Biol.* **20**, 70–80 (2000).
- ten Dam, G. B. *et al.* Regulation of alternative splicing of CD45 by antagonistic effects of SR protein splicing factors. *J. Immunol.* **164**, 5287–5295 (2000).
- Wang, H. Y., Xu, X., Ding, J. H., Bermingham, J. R. Jr & Fu, X. D. SC35 plays a role in T cell development and alternative splicing of CD45. *Mol. Cell* **7**, 331–342 (2001).
- Daoud, R., Da Penha Berzaghi, M., Siedler, F., Hubener, M. & Stamm, S. Activity-dependent regulation of alternative splicing patterns in the rat brain. *Eur. J. Neurosci.* **11**, 788–802 (1999).
- Hofmann, Y., Lorson, C. L., Stamm, S., Androphy, E. J. & Wirth, B. *Htra2-β1* stimulates an exonic splicing enhancer and can restore full-length SMN expression to survival motor neuron 2 (SMN2). *Proc. Natl Acad. Sci. USA* **97**, 9618–9623 (2000).
- Berke, J. D. *et al.* Dopamine and glutamate induce distinct striatal splice forms of Ania-6, an RNA polymerase II-associated cyclin. *Neuron* **32**, 277–287 (2001).
- Xie, J. & Black, D. L. A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels. *Nature* **410**, 936–939 (2001).

59. Black, D. L. Splicing in the inner ear: a familiar tune, but what are the instruments? *Neuron* **20**, 165–168 (1998).
60. Nilsen, T. W. Evolutionary origin of SL-addition trans-splicing: still an enigma. *Trends Genet.* **17**, 678–680 (2001).
61. Blumenthal, T. Trans-splicing and polycistronic transcription in *Caenorhabditis elegans*. *Trends Genet.* **11**, 132–136 (1995).
62. Eul, J., Graessmann, M. & Graessmann, A. Experimental evidence for RNA trans-splicing in mammalian cells. *EMBO J.* **14**, 3226–3235 (1995).
63. Caudevilla, C. *et al.* Heterologous HIV-nef mRNA trans-splicing: a new principle how mammalian cells generate hybrid mRNA and protein molecules. *FEBS Lett.* **507**, 269–279 (2001).
64. Caudevilla, C. *et al.* Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc. Natl Acad. Sci. USA* **95**, 12185–12190 (1998).
65. Takahara, T., Kanazu, S. I., Yanagisawa, S. & Akanuma, H. Heterogeneous Sp1 mRNAs in human HepG2 cells include a product of homotypic trans-splicing. *J. Biol. Chem.* **275**, 38067–38072 (2000).
66. Frantz, S. A. *et al.* Exon repetition in mRNA. *Proc. Natl Acad. Sci. USA* **96**, 5400–5405 (1999).
67. Akopian, A. N. *et al.* Trans-splicing of a voltage-gated sodium channel is regulated by nerve growth factor. *FEBS Lett.* **445**, 177–182 (1999).
68. Caudevilla, C. *et al.* Localization of an exonic splicing enhancer responsible for mammalian natural trans-splicing. *Nucleic Acids Res.* **29**, 3108–3115 (2001).
69. Vellard, M. *et al.* C-myc proto-oncogene: evidence for intermolecular recombination of coding sequences. *Oncogene* **6**, 505–514 (1991).
70. Sullivan, P. M., Petrusz, P., Szpirer, C. & Joseph, D. R. Alternative processing of androgen-binding protein RNA transcripts in fetal rat liver. Identification of a transcript formed by trans splicing. *J. Biol. Chem.* **266**, 143–154 (1991).
71. Li, B. L. *et al.* Human acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) gene organization and evidence that the 4.3-kilobase ACAT-1 mRNA is produced from two different chromosomes. *J. Biol. Chem.* **274**, 11060–11071 (1999).
72. Hirayama, T., Sugino, H. & Yagi, T. Somatic mutations of synaptic cadherin (CNR family) transcripts in the nervous system. *Genes Cells* **6**, 151–164 (2001).
73. Shimizu, A. & Honjo, T. Synthesis and regulation of trans-mRNA encoding the immunoglobulin epsilon heavy chain. *FASEB J.* **7**, 149–154 (1993).
74. Fujieda, S., Lin, Y. Q., Saxon, A. & Zhang, K. Multiple types of chimeric germ-line Ig heavy chain transcripts in human B cells: evidence for trans-splicing of human Ig RNA. *J. Immunol.* **157**, 3450–3459 (1996).
75. Chatterjee, T. K. & Fisher, R. A. Novel alternative splicing and nuclear localization of human RGS12 gene products. *J. Biol. Chem.* **275**, 29660–29671 (2000).
76. Finta, C. & Zaphiropoulos, P. G. Intergenic mRNA molecules resulting from trans-splicing. *J. Biol. Chem.* **277**, 5882–5890 (2002).
77. Labrador, M. *et al.* Protein encoding by both DNA strands. *Nature* **409**, 1000 (2001).
78. Dorn, R., Reuter, G. & Loewendorf, A. Transgene analysis proves mRNA trans-splicing at the complex *mod(mdg4)* locus in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98**, 9724–9729 (2001).
79. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
80. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
81. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
82. Daly, M. Estimating the human gene count. *Cell* **109**, 283–284 (2002).
83. Davidson, E. H. *Genomic Regulatory Systems* (Academic, New York, 2001).
84. Suzuki, Y., Yamashita, R., Nakai, K. & Sugano, S. DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.* **30**, 328–331 (2002).
85. Bono, H., Kasukawa, T., Furuno, M., Hayashizaki, Y. & Okazaki, Y. FANTOM DB: database of functional annotation of RIKEN mouse cDNA clones. *Nucleic Acids Res.* **30**, 116–118 (2002).
86. Kristiansen, T. Z. & Pandey, A. Resources for full-length cDNAs. *Trends Biochem. Sci.* **27**, 266–267 (2002).
87. Sorek, R. & Amitai, M. Piecing together the significance of splicing. *Nature Biotechnol.* **19**, 196 (2001).
88. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nature Genet.* **30**, 13–19 (2002).
89. Hu, G. K. *et al.* Predicting splice variant from DNA chip expression data. *Genome Res.* **11**, 1237–1245 (2001).
90. Shoemaker, D. D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).
91. Yeakly, J. M. *et al.* Profiling alternative splicing on fiber optic arrays. *Nature Biotechnol.* **20**, 1–6 (2002).
92. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. Alternative splicing and genome complexity. *Nature Genet.* **30**, 29–30 (2002).
93. Schmucker, D. *et al.* *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**, 671–684 (2000).
94. Missler, M. & Sudhof, T. C. Neurexins: three genes and 1001 products. *Trends Genet.* **14**, 20–26 (1998).
95. Ptashne, M. & Gann, A. *Genes & Signals* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2002).
96. Reed, R. & Hurt, E. A Conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell* **108**, 523–531 (2002).

Acknowledgements

We thank P. Cramer, B. Graveley, A. Krainer, C. Nabholz and R. Reed for their comments on the manuscript, and R. Hellmiss for the illustrations.

RNA interference

Gregory J. Hannon

Cold Spring Harbour Laboratory, 1 Bungtown Road, Cold Spring Harbour, New York 11724, USA (e-mail: hannon@cshl.org)

A conserved biological response to double-stranded RNA, known variously as RNA interference (RNAi) or post-transcriptional gene silencing, mediates resistance to both endogenous parasitic and exogenous pathogenic nucleic acids, and regulates the expression of protein-coding genes. RNAi has been cultivated as a means to manipulate gene expression experimentally and to probe gene function on a whole-genome scale.

The phenomenon of RNAi was first discovered in the nematode worm *Caenorhabditis elegans* as a response to double-stranded RNA (dsRNA), which resulted in sequence-specific gene silencing¹. Following on from the studies of Guo and Kemphues, who had found that sense RNA was as effective as antisense RNA for suppressing gene expression in worms², Fire, Mello and colleagues¹ were attempting to use antisense RNA as an approach to inhibit gene expression. Their breakthrough was to test the synergy of sense and antisense RNAs, and they duly found that the dsRNA mixture was at least tenfold more potent as a silencing trigger than were sense or antisense RNAs alone¹. Silencing by dsRNAs had a number of remarkable properties — RNAi could be provoked by injection of dsRNA into the *C. elegans* gonad or by introduction of dsRNA through feeding either of dsRNA itself or of bacteria engineered to express it³. Furthermore, exposure of a parental animal to only a few molecules of dsRNA per cell triggered gene silencing throughout the treated animal (systemic silencing) and in its F₁ (first generation) progeny (Fig. 1).

From this discovery emerged the notion that a number of previously characterized, homology-dependent gene-silencing mechanisms might share a common biological root. Several years previously, Richard Jorgensen had been engineering transgenic petunias with the goal of altering pigmentation. But introducing exogenous transgenes did not deepen flower colour as expected. Instead, flowers showed variegated pigmentation, with some lacking pigment altogether (refs 4, 5, and reviewed in ref. 6). This indicated that not only were the transgenes themselves inactive, but also that the added DNA sequences somehow affected expression of the endogenous loci. This phenomenon, called co-suppression, can be produced by highly expressed, single-copy transgenes^{7,8} or by transgenes, expressed at a more modest level, that integrate into the genome in complex, multicopy arrays⁹. In parallel, several laboratories found that plants responded to RNA viruses by targeting viral RNAs for destruction^{10–13}. Notably, silencing of endogenous genes could also be triggered by inclusion of homologous sequences in a virus replicon.

What is clear in retrospect is that both complex transgene arrays and replicating RNA viruses generate dsRNA. In plant systems, dsRNAs that are introduced from exogenous sources or that are transcribed from engineered inverted repeats are potent inducers of gene silencing (reviewed in ref. 14). But co-suppression phenomena are not restricted to plants: similar outcomes have been noted in unicellular organisms, such as *Neurospora*, and in metazoans, such as

Drosophila, *C. elegans* and mammals^{15–18}. In a few cases, silencing has been correlated with integration of transgenes as complex arrays that can produce dsRNA directly, although silencing can also be triggered by the presence of single-copy or dispersed elements¹⁸. What remains a mystery is how, and indeed whether, such elements produce the dsRNA silencing trigger that has become a hallmark of RNAi. It has been proposed that endogenous RNA-directed RNA polymerases (RdRPs) may recognize 'aberrant transcripts' derived from highly expressed loci and convert these into dsRNA¹⁹. Indeed, homologues of these enzymes have proven essential for silencing in *C. elegans*, fungi and plants, and this is discussed below.

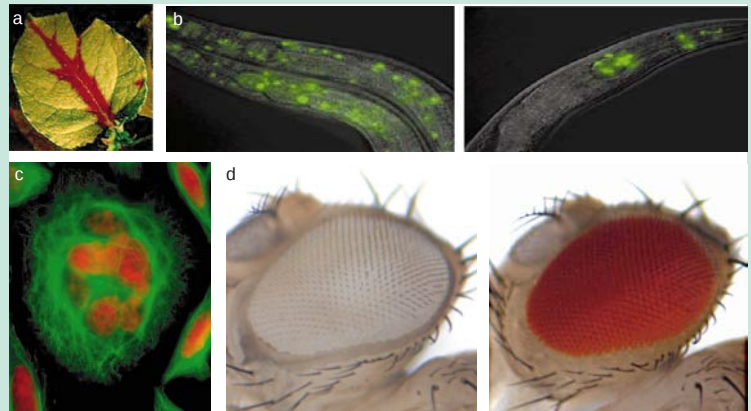
Genetic and biochemical studies have now confirmed that RNAi, co-suppression and virus-induced gene silencing share mechanistic similarities, and that the biological pathways underlying dsRNA-induced gene silencing exist in many, if not most, eukaryotic organisms (Fig. 1). What are the mechanisms by which dsRNAs induce silencing of homologous sequences, either exogenous or endogenous? What are the biological functions of these processes? And how are they related in evolutionarily divergent fungi, plants and animals?

Silencing machinery operates at multiple levels

In *C. elegans*, initial observations were consistent with dsRNA-induced silencing operating at the post-transcriptional level. Exposure to dsRNAs resulted in loss of corresponding messenger RNAs (mRNAs), and promoter and intronic sequences were largely ineffective as silencing triggers¹. A post-transcriptional mode was also consistent with data from plant systems in which exposure to dsRNA²⁰, for example in the form of an RNA virus, triggered depletion of mRNA sequences without an apparent effect on the rate of transcription²¹. Indeed, viral transcripts themselves were targeted, despite the fact that these were synthesized cytoplasmically by transcription of RNA genomes¹⁰. These studies led to the notion that RNAi induced degradation of homologous mRNAs, and this hypothesis has been validated by biochemical analysis.

But the RNAi machinery affects gene expression through additional mechanisms. In plants, exposure to dsRNA induces genomic methylation of sequences homologous to the silencing trigger²². If the trigger shares sequence with a promoter, the targeted gene can become transcriptionally silenced²³. Recent studies have suggested that the RNAi machinery may also affect gene expression at the level of chromatin structure in *Drosophila*, *C. elegans* and fungi (refs 18, 24–26, and R. Martienssen, T. Volpe, I. Hall and S. Grewal, unpublished data). Finally, in *C. elegans*, endogenously

Figure 1 Double-stranded RNA can be introduced experimentally to silence target genes of interest. In plants, silencing can be triggered, for example, by engineered RNA viruses or by inverted repeat transgenes. In worms, silencing can be triggered by injection or feeding of dsRNA. In both of these systems, silencing is systemic and spreads throughout the organism. **a**, A silencing signal moves from the veins into leaf tissue. Green is green fluorescent protein (GFP) fluorescence and red is chlorophyll fluorescence that is seen upon silencing of the GFP transgene. **b**, *C. elegans* engineered to express GFP in nuclei. Animals on the right have been treated with a control dsRNA, whereas those on the left have been exposed to GFP dsRNA. Some neuronal nuclei remain fluorescent, correlating with low expression of a protein required for systemic RNAi⁵⁹. **c**, HeLa cells treated with an ORC6 siRNA and stained for tubulin (green) and DNA (red). Depletion of ORC6 results in accumulation of multinucleated cells. Stable silencing can also be induced by expression of dsRNA as hairpins or snap-back RNAs. **d**, Adult *Drosophila* express a hairpin homologous to the white gene (left), which results in unpigmented eyes compared with wild type (right).



encoded inducers of the RNAi machinery (for example, *lin-4*) operate at the level of protein synthesis²⁷. Although translational control by dsRNA has not been established definitively in other systems, the conservation of *let-7* and related RNAs²⁸ suggests that this regulatory mode may be a further common mechanism through which RNAi pathways control the expression of cellular genes.

Mechanism of post-transcriptional gene silencing

Our present understanding of the mechanisms underlying dsRNA-induced gene silencing is derived from genetic studies in *C. elegans* and plants and from biochemical studies of *Drosophila* extracts. In the latter case, Carthew and colleagues laid the foundations by showing that injection of dsRNA into *Drosophila* embryos induced sequence-specific silencing at the post-transcriptional level²⁹. Sharp and colleagues then tested the possibility that *Drosophila* embryo extracts, previously used to study translational regulation, might be competent for RNAi³⁰. Incubation of dsRNA in these cell-free lysates reduced their ability to synthesize luciferase from a synthetic mRNA. This correlated with destabilization of the mRNA and suggested that dsRNA might bring about silencing by triggering the assembly of a nuclease complex that targets homologous RNAs for degradation.

This effector nuclease, now known as RISC (RNA-induced silencing complex), was isolated from extracts of *Drosophila* S2 cells in which RNAi had been triggered by treatment with dsRNA *in vivo*³¹. A key question was how this complex might identify cognate substrates. Fire and Mello had originally proposed that some derivative of the dsRNA would guide the identification of substrates for RNAi, and the first clue in the hunt for such 'guide RNAs' came from the study of silencing in plants. Hamilton and Baulcombe³² sought antisense RNAs that were homologous to genes being targeted by co-suppression. They found a ~25-nucleotide RNA that appeared only in plant lines containing a suppressed transgene, and found that similar species appeared during virus-induced gene silencing. Similar small RNAs were produced from dsRNAs in *Drosophila* embryo extracts³³, and partial purification of the RISC complex showed that these small RNAs co-fractionated with nuclease activity³¹.

These findings forged a link between transgene co-suppression in plants and RNAi in animals. In addition, a model for RNAi and related silencing phenomenon began to emerge (Fig. 2). According to this model, initiation of silencing occurs upon recognition of dsRNA by a machinery that converts the silencing trigger to ~21–25-nucleotide RNAs. These small interfering RNAs (siRNAs) are a signature of this family of silencing pathways and, by joining an effector complex RISC, they guide that complex to homologous substrates.

This convergence of observations from diverse experimental systems suggested that a conserved biochemical mechanism would lie at the core of homology-dependent gene-silencing responses.

However, the varied biology of dsRNA-induced silencing — for example, the heritable and systemic nature of silencing in *C. elegans* compared to apparently cell-autonomous, non-heritable silencing in *Drosophila* and mammals — suggested that this core machinery probably adapted to meet specific biological needs in different organisms.

The initiation step

The model outlined in Fig. 2 implies that the dsRNA silencing trigger is cleaved to produce siRNAs. Support for this emerged first from studies of *Drosophila* embryo extracts, which contained an activity capable of processing long dsRNA substrates into ~22-nucleotide fragments³³. Analysis of these RNAs showed that they were double stranded and contained 5'-phosphorylated termini^{33,34}. The quest for the enzyme that initiates RNAi led to the RNase III ribonuclease family, which displays specificity for dsRNAs and generates such termini.

RNase III enzymes can be divided into three classes based upon domain structure: bacterial RNase III contains a single catalytic domain and a dsRNA-binding domain; Drosha family nucleases contain dual catalytic domains³⁵; and a third family also contains dual catalytic domains and additional helicase and PAZ motifs³⁶. Members of this third class of RNases were found to process dsRNA into siRNAs and were therefore proposed to initiate RNAi³⁶. This family, now named the Dicer enzymes, are evolutionarily conserved, and proteins from *Drosophila*, *Arabidopsis*, the insect *Spodoptera frugiperda*, tobacco, *C. elegans*, mammals and *Neurospora* have all been shown to recognize and process dsRNA into siRNAs of a characteristic size for the relevant species (refs 36, 37, and A. M. Denli and G.J.H., unpublished data). Genetic evidence has also emerged from *C. elegans* and *Arabidopsis* that is consistent with Dicer acting in the RNAi pathway: Dicer is required for RNAi in the *C. elegans* germline^{37–39}, and a hypomorphic allele of *Carpel Factory* can intensify the phenotypes of weak *Argonaute-1* alleles in *Arabidopsis* (C. Kidner and R. Martienssen, personal communication).

Recently, the structure of an RNase III catalytic domain has led to a model for the generation of ~22-nucleotide RNAs by Dicer cleavage⁴⁰ (Fig. 2). It is thought that bacterial RNase III functions as a dimeric enzyme and, in the structural model, antiparallel RNase III domains produce two compound catalytic centres, each of which is formed by contributions from both monomers. The sequences of Dicer and Drosha RNase III domains reveal deviations from the consensus in both enzymes. Introduction of these alterations into bacterial RNase III permitted a genetic test for domain function: defects were noted upon introduction of residues that form part of the catalytic centre from the second RNase III domain of Dicer family members. Antiparallel alignment of Dicer's RNase III motifs on a dsRNA substrate could produce four compound active sites, but the central two of these would be inactive. In this way, cleavage would

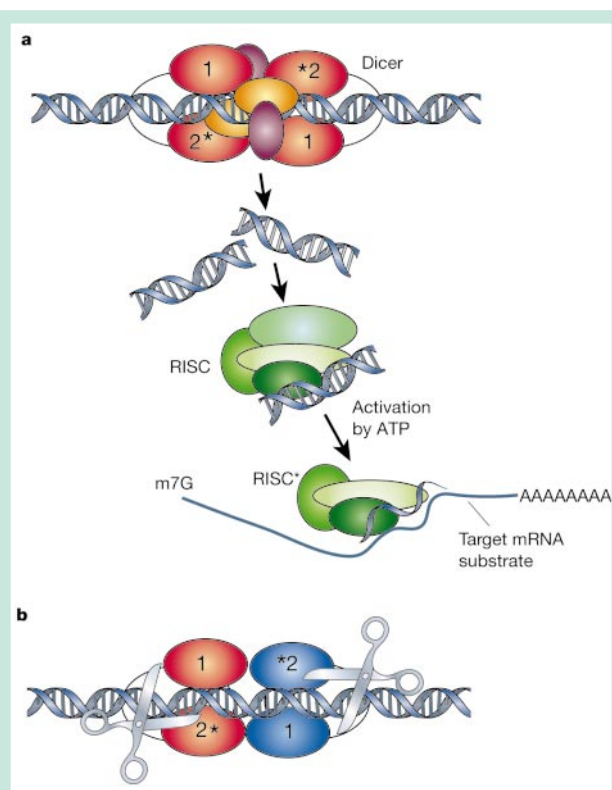


Figure 2 Dicer and RISC (RNA-induced silencing complex). **a**, RNAi is initiated by the Dicer enzyme (two Dicer molecules with five domains each are shown), which processes double-stranded RNA into ~22-nucleotide small interfering RNAs³⁶. Based upon the known mechanisms for the RNase III family of enzymes, Dicer is thought to work as a dimeric enzyme. Cleavage into precisely sized fragments is determined by the fact that one of the active sites in each Dicer protein is defective (indicated by an asterisk), shifting the periodicity of cleavage from ~9–11 nucleotides for bacterial RNase III to ~22 nucleotides for Dicer family members⁴⁰. The siRNAs are incorporated into a multicomponent nuclease, RISC (green). Recent reports suggest that RISC must be activated from a latent form, containing a double-stranded siRNA to an active form, RISC*, by unwinding of siRNAs⁴¹. RISC* then uses the unwound siRNA as a guide to substrate selection³¹. **b**, Diagrammatic representation of Dicer binding and cleaving dsRNA (for clarity, not all the Dicer domains are shown, and the two separate Dicer molecules are coloured differently). Deviations from the consensus RNase III active site in the second RNase III domain inactivate the central catalytic sites, resulting in cleavage at 22-nucleotide intervals.

occur at ~22-base intervals, and subtle alterations in Dicer structure could alter the spacing of these catalytic centres and explain the species-specific variation in siRNA length (A. Denli and G.J.H, unpublished results).

The effector step

In the *Drosophila* system, RNAi is enforced by RISC, a protein–RNA effector nuclease complex that recognizes and destroys target mRNAs. The first subunit of RISC to be identified was the siRNA, which presumably identifies substrates through Watson–Crick base-pairing³¹. Zamore and colleagues have recently shown that RISC is formed in embryo extracts as a precursor complex of ~250K; this becomes activated upon addition of ATP to form a ~100K complex that can cleave substrate mRNAs⁴¹. Cleavage is apparently endonucleolytic, and occurs only in the region homologous to the siRNA. siRNAs are double-stranded duplexes with two-nucleotide 3' overhangs and 5'-phosphate termini^{33,34}, and this configuration is functionally important for incorporation into RISC complexes^{34,41}. However, single-stranded siRNAs should be most effective at seeking

homologous targets, and one intriguing correlation with the transition of RISC zymogens to active enzymes is siRNA unwinding⁴¹.

My laboratory has purified RISC from *Drosophila* S2 cells as a ~500K ribonucleoprotein with slightly different characteristics^{31,42}. In embryo extracts, RISC* (the 100K active RISC species) cleaves its substrates endonucleolytically⁴¹. Intermediate cleavage products are never observed in even the most highly purified RISC preparations from S2 cells, suggesting the presence of an exonuclease in this enzyme complex. Therefore, the complex formed *in vivo* probably contains additional factors that account for observed differences in size and activity. Alternatively, RISC purified from S2 cells may become activated — perhaps changing size and subunit composition — upon incubation with ATP.

RISC from S2 cells co-purifies with AGO2, a member of the *Argonaute* gene family⁴². Argonaute proteins were first identified in *Arabidopsis* mutants that produced altered leaf morphology⁴³, and form a large, evolutionarily conserved gene family with representatives in most eukaryotic genomes, with the possible exception of *Saccharomyces cerevisiae* (reviewed in ref. 44). These proteins are characterized by the presence of two homology regions, the PAZ domain and the Piwi domain, the latter being unique to this group of proteins. The PAZ domain also appears in Dicer proteins, and may be important in the assembly of silencing complexes³⁶.

Argonaute proteins were linked to RNAi by genetic studies in *C. elegans*, whose genome contains >20 related genes. The *rde-1* gene was isolated by Mello and colleagues²⁵ from a mutant worm that was unable to sustain RNAi in germline or soma. Using genetic methods, Grishok and colleagues⁴⁵ found a requirement for RDE-1 and RDE-4 for initiation of silencing in a parental animal; however, neither function was required for systemic silencing in F₁ progeny. In contrast MUT-7 (ref. 46) and RDE-2 were both dispensable in the parent, but were required in their progeny.

Rationalizing these results with the simple model proposed above is difficult. Indeed, RDE-4 is a small dsRNA-binding protein, and both RDE-1 and RDE-4 can interact with *C. elegans* Dicer (H. Tabara *et al.*, unpublished data). Perhaps RDE-4 initially recognizes dsRNA and delivers it to the Dicer enzyme. This would be consistent with the observation that siRNA levels are greatly reduced in worms that lack RDE-4 function, but are abundant in worms that lack RDE-1 (ref. 47). Similarly, in *Neurospora*, mutations in the Argonaute family member *qde-2* eliminate quelling (transgene co-suppression), but do not alter accumulation of siRNAs⁴⁸. Thus RDE-1, and perhaps other Argonaute proteins as well, might shuttle siRNAs to appropriate effector complexes (RISCs). Consistent with this notion, we have detected transient interactions in S2 cell extracts between Dicer and Argonaute family members (ref. 42, and A. Caudy, unpublished data). This model has implications for signal amplification and systemic silencing.

Amplification and spreading of silencing

One of the most provocative aspects of RNAi in *C. elegans* is its ability to spread throughout the organism, even when triggered by minute quantities of dsRNA¹. Similar systemic silencing phenomena have been observed in plants, in which silencing could pervade a plant or even be transferred to a naive grafted scion⁴⁹. Accounting for these phenomena requires firstly a system to pass a signal from cell to cell, and secondly a strategy for amplifying the signal.

Recently, a phenomenon termed 'transitive RNAi' has provided some useful clues. Transitive RNAi refers to the movement of the silencing signal along a particular gene (Fig. 3). For example, in *C. elegans*, targeting the 3' portion of a transcript results in suppression of that mRNA and in the production of siRNAs homologous to the targeted region. In addition, siRNAs complementary to regions of the transcript upstream from the area targeted directly by the silencing trigger also appear and accumulate⁵⁰. If these siRNAs are complementary to other RNAs, those are also targeted (hence, 'transitive' RNAi).

In both plants and *C. elegans*, dsRNA-induced silencing requires proteins similar in sequence to a tomato RNA-directed RNA

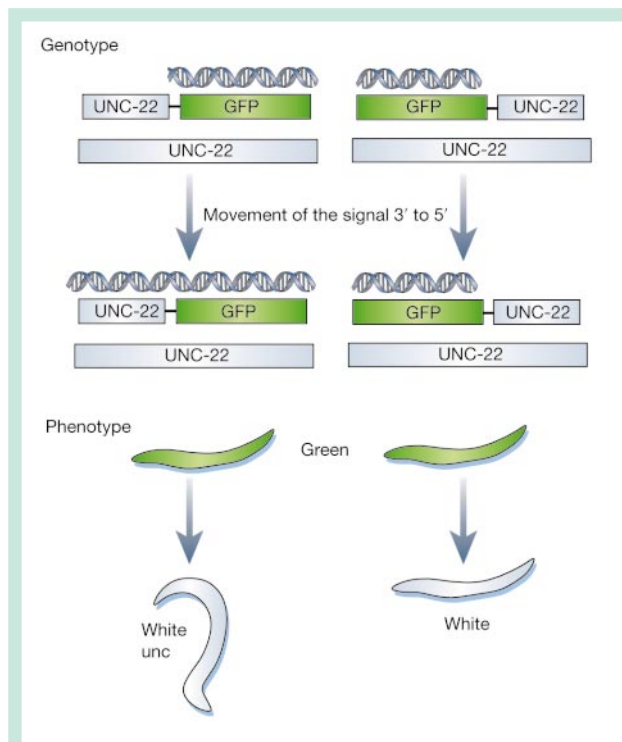


Figure 3 Transitive RNAi. In transitive RNAi in *C. elegans*, silencing can travel in a 3' to 5' direction on a specific mRNA target⁵⁰. The simplest demonstration comes from the creation of fusion transcripts. Consider a fragment of green fluorescent protein (GFP) fused 3' to a segment of UNC-22 (left). Targeting GFP abolishes fluorescence but also creates an unexpected, uncoordinated phenotype. This occurs because of the production of double-stranded RNA and consequently small interfering RNAs homologous to the endogenous UNC-22 gene. In a case in which GFP is fused 5' to the UNC-22 fragment (right), GFP dsRNA still ablates fluorescence but does not produce an uncoordinated phenotype.

polymerase (RdRP)⁵¹, which could be involved in amplifying the RNAi signal. However, only the tomato enzyme has been shown to possess polymerase activity, and biochemical studies will be required to establish definitively the role these proteins play in RNAi. In *Arabidopsis*, SDE1/SGS2 is required for transgene silencing, but not for virally induced gene silencing (VIGS)^{19,52}. This suggests that SDE1/SGS2 may act as an RdRP, as viral replicases could substitute for this function in VIGS. In *Neurospora*, QDE-1 is required for efficient quelling⁵³. EGO-1 is essential for RNAi in the germline of *C. elegans*⁵⁴, and another RdRP homologue, RRF-1/RDE-9, is required for silencing in the soma⁵⁰ (D. Conte and C. Mello, unpublished data).

These genetic studies have led to a model for transitive RNAi in which siRNAs might prime the synthesis of additional dsRNA by RdRPs. RdRP activity has been reported recently from *Drosophila* embryo extracts⁵⁵, although transitive RNAi has yet to be observed in flies. While numerous experiments suggest that an RdRP is not required for RNAi in *Drosophila* extracts, the possibility remains that such an enzyme might act, for example, in triggering RNAi by the production of dsRNA from dispersed, multicopy transgenes.

The fact that RDE-1 and RDE-4 are required only for initiation of RNAi in parental *C. elegans* adds an additional layer of complexity to the model. Perhaps exogenous dsRNAs are recognized initially in manner that is distinct from recognition of secondary dsRNA, which may be produced by RdRPs. For example, the proposed function of RDE-4 in delivering dsRNA to Dicer could be substituted for secondary dsRNAs by another hypothetical protein. Alternatively, Dicer could exist in a stable complex with an RdRP, making dsRNA delivery unnecessary. The requirement for RRF-1/RDE-9 throughout the

C. elegans soma — and the similar requirement for SDE1/SGS2 in plants — also suggests that most RNAi in these systems is driven by secondary siRNAs produced through the action of RdRPs.

However, other possibilities also exist. Indeed, in plants, transitive RNAi travels in both 3'→5' and 5'→3' directions⁵⁶, which is inconsistent with the simple notion of siRNAs priming dsRNA synthesis. Instead, one can imagine that genomic loci may serve as a reservoir for silencing. In some systems, it is known that exposure to dsRNA can produce alterations in chromatin structure, which could lead to the production of 'aberrant' mRNAs that are substrates for conversion to dsRNA by RdRPs. This model would permit bi-directional spread, as such an expansion of altered chromatin structure is an established phenomenon. Moreover, a similar model could explain co-suppression that is occasionally triggered by single-copy, dispersed transgenes. Finally, this model would be consistent with transitive effects that have been observed for both transcriptional and post-transcriptional silencing in *Drosophila*, which operate in the absence of any homology in the transcribed RNA, and thus differ from 'transitive RNAi' in *C. elegans*^{18,24}. But support for a genome-based amplification model remains elusive, as does the nature of the 'aberrant' RNAs that trigger siRNA formation and an explanation for how chromatin modifications could induce their production.

Although these models suggest mechanisms for cell-autonomous amplification of the silencing signal, the character of the signal that transmits systemic silencing in plants and animals is unknown. Two candidates are siRNAs themselves or long dsRNAs, perhaps formed via RdRP-dependent amplification.

Note that, in plants, two types of transmission must be considered. The first is short-range, cell-to-cell transmission. Plant cells are intimately connected through cytoplasmic bridges known as plasmodesmata. Movement of RNA and proteins via these cell-cell junctions is well known, and it is likely that either long dsRNA or siRNAs could be passed through these connections. But the silencing signal must also be passed over a longer range through the plant vasculature⁵⁷. In this regard, studies of a viral silencing inhibitor have provided evidence against siRNAs being critical for systemic silencing in plants. Hc-Pro suppresses silencing and also interferes with the production of siRNAs from dsRNA triggers⁵⁸. Expression of Hc-Pro does not interfere with transgene methylation, which results in transcriptional gene silencing (TGS) if present in the promoter and which may contribute to post-transcriptional gene silencing (PTGS) if present in the transcribed sequence. Hc-Pro expression in a silenced rootstock relieves silencing and inhibits siRNA production, but a systemic signal can still be passed from this rootstock to an engrafted scion lacking Hc-Pro expression.

Recently, Hunter and colleagues identified a protein in *C. elegans* that is required for systemic silencing⁵⁹. The *sid-1* gene encodes a transmembrane protein that may act as a channel for import of the silencing signal. Expression of *sid-1* is largely lacking from neuronal cells, perhaps explaining initial observations that *C. elegans* neurons were resistant to systemic RNAi. SID-1 homologues are absent from *Drosophila*, consistent with a lack of systemic transmission of silencing in flies, but are present in mammals, raising the possibility that some aspects of RNAi may act non-cell autonomously in mammals.

Other components of the RNAi machinery

A combination of genetics and biochemistry has led to much progress towards understanding the mechanism of PTGS, but many questions remain. In *Drosophila* embryo extracts, pre-RISC becomes activated upon unwinding of siRNAs in an ATP-dependent process. A number of different helicases have been identified in searches for RNAi-deficient mutants (for example, QDE-3, MUT6 and MUT-14), and any of these might be candidates for a RISC activator^{60–62}. Additionally, the identities of RISC-associated nucleases that cleave targeted mRNAs remain elusive. Studies of RISC formed in embryo extracts suggest an endonuclease that cleaves the siRNA–mRNA hybrid near the middle of the duplex, while RISC

formed *in vivo* may have additional exonuclease activities. The MUT-7 protein, which is essential for RNAi in the *C. elegans* germ line, has nuclease homology, but a *Drosophila* relative of this protein has not yet been found in RISC (ref. 46, and S. Hammond, unpublished data). The efficiency of RNAi suggests an active mechanism for searching the transcriptome for homologous substrates. Most *Drosophila* RISC might be associated with the ribosome³¹, and recent studies have extended this observation to trypanosomes (*E. Ullu*, unpublished data). Finally, relationships between the RNAi machinery and other aspects of RNA metabolism in the cell must be explored. For example, genetic evidence⁶³ suggests a link between RNAi and nonsense-mediated decay, raising the possibility that the RNAi machinery may be important in destruction of improperly processed mRNAs or in the general regulation of mRNA stability.

RNAi and the genome

In plants, dsRNA induces genomic methylation at sites of sequence homology (ref. 22, reviewed in ref. 64). Methylation is asymmetric and is not restricted to CpG or CpXpG sequences. If methylation occurs in the coding sequence, it has no apparent effect on the transcription of the locus, although silencing still occurs at the post-transcriptional level. Methylation of the promoter sequence induces TGS²³, which unlike PTGS is stable and heritable²¹. Thus, dsRNA can clearly trigger alterations at the genomic level, but the degree to which these alterations are relevant to PTGS remains uncertain.

Recent studies have begun to generalize the notion of an intimate connection between the RNAi machinery and the genome, and to draw mechanistic links between PTGS and TGS. For example, in *C. elegans*, *mut-7* and *rde-2* mutations de-repress transgenes that are silenced at the level of transcription by a polycomb-dependent mechanism²⁵. Polycomb-group proteins function by organizing chromatin into 'open' or 'closed' conformations, creating stable and heritable patterns of gene expression. Recently, Goldstein and colleagues found that the polycomb proteins MES-3, MES-4 and MES-6 are required for RNAi, at least under some experimental conditions²⁶. Mutant worms were deficient in the RNAi response if high levels of dsRNA were injected, but were not deficient in the presence of limiting dsRNA. Of course, the effects of these mutants could be indirect, altering the expression of other elements or regulators of the RNAi pathway. However, links between altered chromatin structures and dsRNA-induced gene silencing have also emerged from plant and *Drosophila* systems. In particular, alterations of either methyltransferases (*MET1*) or chromatin remodelling complexes (for example, *DDM1*) can affect both the degree and persistence of silencing in *Arabidopsis*^{21,65}. Conversely, mutations in genes required

for PTGS (for example, *AGO1* and *SGS2*) decrease both co-suppression and transgene methylation⁶⁶. Furthermore, mutation of *piwi*, a relative of the RISC component *Argonaute-2*, compromises co-suppression of dispersed transgenes in *Drosophila* at both the post-transcriptional and transcriptional levels²⁴.

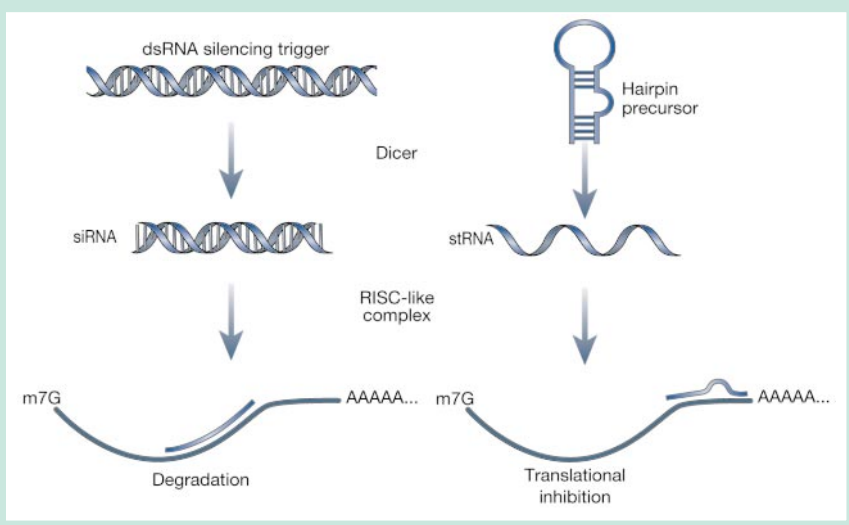
Thus, one of the most fascinating and least-explored responses to dsRNA involves a possible recognition of genomic DNA by derivatives of the silencing trigger, possibly siRNAs. One model suggests that a variant, nuclear RISC carries a chromatin remodelling complex rather than a ribonuclease to its cognate target. Indeed, Martienssen, Grewal and colleagues have recently noted a requirement for relatives of Dicer and RISC components in the silencing of centromeric repeats in *Schizosaccharomyces pombe* (T. Volpe, C. Kidner, I. Hall, S. Grewal and R. Martienssen, personal communication). It seems therefore that a principal biological function of the RNAi machinery may be to form heterochromatic domains in the nucleus that are critical for genome organization and stability.

Biological functions of RNAi

Because target identification depends upon Watson–Crick base-pairing interactions, the RNAi machinery can be both flexible and exquisitely specific. Thus, this regulatory paradigm may have been adapted and adopted for numerous cellular functions. For example, in plants, RNAi forms the basis of VIGS, suggesting an important role in pathogen resistance. An elegant proof of this hypothesis comes from the genetic links between virulence and RNAi pathways (refs 52, 67, and reviewed in ref. 68). Many plant viruses encode suppressors of PTGS that are essential for pathogenesis, and these virulence determinants can be masked by host mutations in silencing pathways. RNAi has also been linked to the control of endogenous parasitic nucleic acids. In *C. elegans*, some RNAi-deficient strains are also 'mutators' owing to increased mobility of endogenous transposons^{25,46}. In many systems, transposons are silenced by their packaging into heterochromatin (reviewed in ref. 64). Therefore, it is tempting to speculate that RNAi may stabilize the genome by sequestering repetitive sequences such as mobile genetic elements, preventing transposition and making repetitive elements unavailable for recombination events that would lead to chromosomal translocations. However, it remains to be determined whether RNAi regulates transposons through effects at the genomic level or by post-transcriptionally targeting mRNAs (for example, those encoding transposases) that are required for transposition.

A role for RNAi pathways in the normal regulation of endogenous protein-coding genes was originally suggested through the analysis of plants and animals containing dysfunctional RNAi components.

Figure 4 Small interfering RNAs versus small temporal RNAs. Double-stranded siRNAs of length ~21–23 nucleotides are produced by Dicer from dsRNA silencing triggers. Characteristic of RNase III products, these have two-nucleotide 3' overhangs and 5'-phosphorylated termini. To trigger target degradation with maximum efficiency, siRNAs must have perfect complementarity to their mRNA target (with the exception of the two terminal nucleotides, which contribute only marginally to recognition). stRNAs, such as *lin-4* and *let-7*, are transcribed from the genome as hairpin precursors. These are also processed by Dicer, but in this case, only one strand accumulates. Notably, neither *lin-4* nor *let-7* show perfect complementarity to their targets. In addition, stRNAs regulate targets at the level of translation rather than RNA degradation. It remains unclear whether the difference in regulatory mode results from a difference in substrate recognition or from incorporation of siRNAs and stRNAs into distinct regulatory complexes.



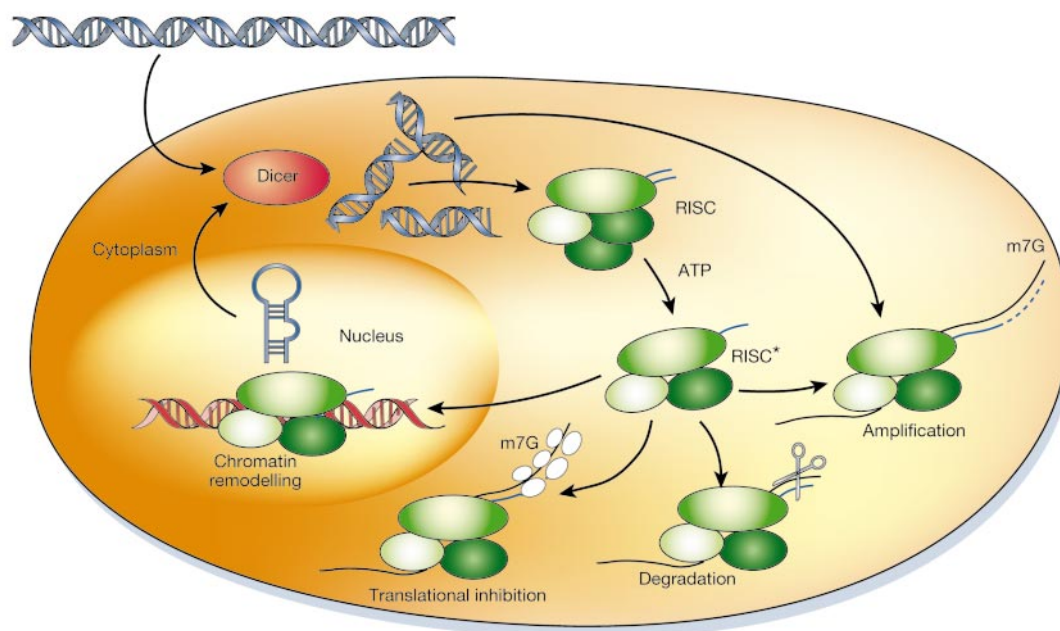


Figure 5 A model for the mechanism of RNAi. Silencing triggers in the form of double-stranded RNA may be presented in the cell as synthetic RNAs, replicating viruses or may be transcribed from nuclear genes. These are recognized and processed into small interfering RNAs by Dicer. The duplex siRNAs are passed to RISC (RNA-induced silencing complex), and the complex becomes activated by unwinding of the duplex. Activated RISC complexes can regulate gene expression at many levels. Almost

certainly, such complexes act by promoting RNA degradation and translational inhibition. However, similar complexes probably also target chromatin remodelling. Amplification of the silencing signal in plants may be accomplished by siRNAs priming RNA-directed RNA polymerase (RdRP)-dependent synthesis of new dsRNA. This could be accomplished by RISC-mediated delivery of an RdRP or by incorporation of the siRNA into a distinct, RdRP-containing complex.

Mutations in the *Argonaute-1* gene of *Arabidopsis*, for example, cause pleiotropic developmental abnormalities that are consistent with alterations in stem-cell fate determination⁴³. A hypomorphic mutation in *Carpel Factory*, an *Arabidopsis* Dicer homologue, causes defects in leaf development and overproliferation of floral meristems⁶⁹. Mutations in *Argonaute* family members in *Drosophila* also impact normal development. In particular, mutations in *Argonaute-1* have drastic effects on neuronal development⁷⁰, and *piwi* mutants have defects in both germline stem-cell proliferation and maintenance⁷¹.

This should not be interpreted as a demonstration that PTGS pathways regulate endogenous gene expression *per se*. In fact, separation-of-function *ago1* mutants have recently been isolated that preferentially affect PTGS⁷² without affecting development. Mutations in *Zwille*, another *Argonaute* family member, also alter stem-cell maintenance⁷³, and this occurs without perceptible impact on dsRNA-mediated silencing⁷². Thus, components of the RNAi machinery, and related gene products, may function in related but separable pathways of gene regulation.

A possible mechanism underlying the regulation of endogenous genes by the RNAi machinery emerged from the study of *C. elegans* containing mutations in their single Dicer gene, *DCR-1*. Unlike most other RNAi-deficient worm mutants, *dcr-1* animals were neither normal nor fertile: the mutation induced a number of phenotypic alterations in addition to its effect on RNAi^{37–39,74}. Intriguingly, Dicer mutants showed alterations in developmental timing similar to those observed in *let-7* and *lin-4* mutants. The *lin-4* gene was originally identified as a mutant that affects larval transitions⁷⁵, and *let-7* was subsequently isolated as a similar heterochronic mutant²⁸. These loci encode small RNAs, which are synthesized as ~70-nucleotide precursors and post-transcriptionally processed to a ~21-nucleotide mature form. Genetic and biochemical studies have indicated that these RNAs are processed by Dicer^{37–39,74}.

The small temporal RNAs (stRNAs) encoded by *let-7* and *lin-4* are negative regulators of specific protein-coding genes, as might be expected if stRNAs trigger RNAi. However, stRNAs do not trigger mRNA degradation, but regulate expression at the translational level^{76,77}. This raised the possibility that stRNAs and RNAi might be linked only by the processing enzyme Dicer. However, Mello and colleagues demonstrated a requirement for Argonaute family proteins (that is, Alg-1 and Alg-2) in both stRNA biogenesis and stRNA-mediated suppression³⁹, which led to a model in which the effector complexes containing siRNAs and stRNAs are closely related, but regulate expression by distinct mechanisms (Fig. 4). Neither LIN-4 nor LET-7 forms a perfect duplex with its cognate target⁷⁸. Thus, in one possible model an analogous RISC complex is formed containing either siRNAs or stRNAs. In the former case, cleavage is dependent upon perfect complementarity, while in the latter, cleavage does not occur, but the complex blocks ribosomal elongation. Alternatively, siRNAs and stRNAs may be discriminated and enter related but distinct complexes that target substrates for degradation or translational regulation, respectively. Consistent with this latter model is the observation that siRNAs or exogenously supplied hairpin RNAs that contain single mismatches with their substrates fail to repress, rather than simply shifting their regulatory mode to translational inhibition^{34,79,80}.

In this scenario, RISC may be viewed as a flexible platform upon which different regulatory modules may be superimposed (Fig. 5). The core complex would be responsible for receiving the small RNA from Dicer and using this as a guide to identify its homologous substrate. Depending upon the signal (for example, its structure and localization), different effector functions could join the core: in RNAi, nucleases would be incorporated into RISC, whereas in stRNA-mediated regulation, translational repressors would join the complex. Transcriptional silencing could be accomplished by the inclusion of chromatin remodelling factors, and one could imagine other adaptations might exist.

Whether or not RISC is a flexible regulator becomes particularly important in light of recent findings that *let-7* and *lin-4* are archetypes of a large class of endogenously encoded small RNAs. Over 100 of these microRNAs or miRNAs have now been identified in *Drosophila*, *C. elegans* and mammals^{81–84}, and although their functions are unknown, their prevalence hints that RNAi-related mechanisms may have pervasive roles in controlling gene expression. In this regard, a number of miRNAs from *Drosophila* are partially complementary to two sequences, the K box and the Brd box, that mediate post-transcriptional regulation of numerous mRNAs⁸⁵.

RNAi and genomics

RNAi has evolved into a powerful tool for probing gene function. In *C. elegans*, testing the functions of individual genes by RNAi has now extended to analysis of nearly all of the worm's predicted ~19,000 genes (J. Ahringer, unpublished data). Similar strategies are being pursued in other organisms, including plants (D. Baulcombe and P. Waterhouse, personal communication). Although it seemed for some time that deploying RNAi in mammalian systems would not be feasible, the first hint that the technology might work came when RNAi was demonstrated in early mouse embryos^{86,87}. But this appeared to be of limited utility, as mammalian somatic cells, but not some embryonic cells, exhibit nonspecific responses to dsRNA which would obscure sequence-specific silencing. One of these is the RNA-dependent protein kinase (PKR) pathway, which responds to dsRNA by phosphorylating EIF-2 α and nonspecifically arresting translation⁸⁸. Tuschl and colleagues then showed that siRNAs themselves could be used to induce effective silencing in many mammalian cells⁷⁹. These small RNAs, which are chemically synthesized mimics of Dicer products, are presumably incorporated into RISC and target cognate substrates for degradation. The siRNAs are too small to induce nonspecific dsRNA responses such as PKR⁸⁹.

One drawback that siRNAs have is that their effects are transient, as mammals apparently lack the mechanisms that amplify silencing in worms and plants. In several systems, including plants, *Drosophila*, *C. elegans* and trypanosomes, RNAi has been made stable and heritable by enforced expression of the silencing trigger, usually as an inverted repeat sequence forming a hairpin structure *in vivo*^{90–95}. We have reported mammalian cell lines in which genes are stably suppressed by RNAi through the expression of a 500-base-pair dsRNA⁹⁶. However, this approach was limited to cell types that lacked generic responses to dsRNA such as the PKR pathway. Recently, we and others have shown that short hairpin RNAs (shRNAs) modelled on miRNAs can be used to manipulate gene expression experimentally^{80,97,98}. These may be expressed *in vivo* from RNA polymerase III (Pol III) promoters to induce stable suppression in mammalian cells.

The availability of stable triggers of RNAi builds upon the utility of siRNAs in several ways. Induced phenotypes can now be observed over long time spans. Stably engineered cells can be assayed either *in vitro* or *in vivo*, perhaps testing the angiogenic or metastatic potential of tumour cells in xenograft models. RNAi may potentially be used to create hypomorphic alleles rapidly in transgenic mice. If inducible Pol III promoters were used^{99,100}, this could permit a powerful approach akin to the use of tissue-specific Gal4-drivers in *Drosophila*. Finally, shRNAs could be combined with existing high-efficiency gene delivery vehicles to create *bona fide* RNAi-based therapeutics. In this regard, we have successfully delivered shRNAs from replication-deficient retroviruses, and foresee numerous applications for *ex vivo* manipulation of stem cells based upon this paradigm. For example, a patient's own bone marrow stem cells could be engineered to resist HIV infection by targeting either the HIV RNA itself or receptors necessary for HIV infection (for example, CCR5). Furthermore, we see no conceptual barrier to incorporating this strategy for targeted suppression into adenovirus or herpesvirus-based delivery vehicles. Ultimately, the exquisite specificity of RNAi may make it possible

to silence a disease-causing mutant allele specifically, such as an activated oncogene, without affecting the normal allele.

Perspective

Over the past few years, the way in which cells respond to dsRNA by silencing homologous genes has revealed a new regulatory paradigm in biology. This response can be triggered in many different ways, ranging from experimental introduction of synthetic silencing triggers to the transcription of endogenous RNAs that regulate gene expression. We are only beginning to appreciate the mechanistic complexity of this process and its biological ramifications. RNAi has already begun to revolutionize experimental biology in organisms ranging from unicellular protozoans to mammals. RNAi has been applied on the whole-genome scale in *C. elegans* and this goal is being pursued in plant systems. My laboratory, as part of the larger cancer genomics effort, has undertaken to target, individually, every gene in the human genome using expressed shRNAs. This will permit large-scale loss-of function genetic screens and rapid tests for genetic interactions to be performed for the first time in mammalian cells. Such approaches hold tremendous promise for unleashing the dormant potential of sequenced genomes. □

1. Fire, A. et al. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
2. Guo, S. & Kemphues, K. J. *par-1*, a gene required for establishing polarity in *C. elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell* **81**, 611–620 (1995).
3. Timmons, L. & Fire, A. Specific interference by ingested dsRNA. *Nature* **395**, 854 (1998).
4. van der Krol, A. R., Mur, L. A., de Lange, P., Mol, J. N. & Stuitje, A. R. Inhibition of flower pigmentation by antisense CHS genes: promoter and minimal sequence requirements for the antisense effect. *Plant Mol. Biol.* **14**, 457–466 (1990).
5. Napoli, C. A., Lemieux, C., & Jorgensen, R. Introduction of a chimeric chalcone synthase gene in *Petunia* results in reversible cosuppression of homologous genes *in trans*. *Plant Cell* **2**, 279–289 (1990).
6. Jorgensen, R. Altered gene expression in plants due to trans interactions between homologous genes. *Trends Biotechnol.* **8**, 340–344 (1990).
7. Jorgensen, R. A., Cluster, P. D., English, J., Que, Q. & Napoli, C. A. Chalcone synthase cosuppression phenotypes in *Petunia* flowers: comparison of sense vs. antisense constructs and single-copy vs. complex T-DNA sequences. *Plant Mol. Biol.* **31**, 957–973 (1996).
8. Elmayan, T. & Vaucheret, H. Single copies of a strongly expressed 35S-driven transgene undergo post-transcriptional silencing. *Plant J.* **9**, 787–797 (1996).
9. Que, Q., Wang, H. Y., English, J. & Jorgensen, R. The frequency and degree of cosuppression by sense chalcone synthase transgenes are dependent on promoter strength and are reduced by premature nonsense codons in the transgene coding sequence. *Plant Cell* **9**, 1357–1368 (1997).
10. Ruiz, M. T., Voinnet, O. & Baulcombe, D. C. Initiation and maintenance of virus-induced gene silencing. *Plant Cell* **10**, 937–946 (1998).
11. Angell, S. M. & Baulcombe, D. C. Consistent gene silencing in transgenic plants expressing a replicating potato virus X RNA. *EMBO J.* **16**, 3675–3684 (1997).
12. Dougherty, W. G. et al. RNA-mediated virus resistance in transgenic plants: exploitation of a cellular pathway possibly involved in RNA degradation. *Mol. Plant Microbe Interact.* **7**, 544–552 (1994).
13. Kumagai, M. H. et al. Cytoplasmic inhibition of carotenoid biosynthesis with virus-derived RNA. *Proc. Natl Acad. Sci. USA* **92**, 1679–1683 (1995).
14. Bernstein, E., Denli, A. M. & Hannon, G. J. The rest is silence. *RNA* **7**, 1509–1521 (2001).
15. Romano, N. & Macino, G. Quelling: transient inactivation of gene expression in *Neurospora crassa* by transformation with homologous sequences. *Mol. Microbiol.* **6**, 3343–3353 (1992).
16. Fire, A., Albertson, D., Harrison, S. W. & Moerman, D. G. Production of antisense RNA leads to effective and specific inhibition of gene expression in *C. elegans* muscle. *Development* **113**, 503–514 (1991).
17. Dernburg, A. F., Zaslavsky, J., Colaiaicovo, M. P. & Villeneuve, A. M. Transgene-mediated cosuppression in the *C. elegans* germ line. *Genes Dev.* **14**, 1578–1583 (2000).
18. Pal-Bhadra, M., Bhadra, U. & Birchler, J. A. Cosuppression in *Drosophila*: gene silencing of *Alcohol dehydrogenase* by white-*Adh* transgenes is *Polycomb* dependent. *Cell* **90**, 479–490 (1997).
19. Dalmay, T., Hamilton, A., Rudd, S., Angell, S. & Baulcombe, D. C. An RNA-dependent RNA polymerase gene in *Arabidopsis* is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. *Cell* **101**, 543–553 (2000).
20. de Carvalho, F. et al. Suppression of beta-1,3-glucanase transgene expression in homozygous plants. *EMBO J.* **11**, 2595–2602 (1992).
21. Jones, L., Ratcliff, F. & Baulcombe, D. C. RNA-directed transcriptional gene silencing in plants can be inherited independently of the RNA trigger and requires Met1 for maintenance. *Curr. Biol.* **11**, 747–757 (2001).
22. Wassenaar, M., Heimes, S., Riedel, L. & Sanger, H. L. RNA-directed de novo methylation of genomic sequences in plants. *Cell* **76**, 567–576 (1994).
23. Mette, M. F., Aufsatz, W., van der Winden, J., Matzke, M. A. & Matzke, A. J. Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO J.* **19**, 5194–5201 (2000).
24. Pal-Bhadra, M., Bhadra, U. & Birchler, J. A. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in *Drosophila*. *Mol Cell* **9**, 315–327 (2002).
25. Tabara, H. et al. The *rde-1* gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99**, 123–132 (1999).
26. Dudley, N. R., Labbe, J. C. & Goldstein, B. Using RNA interference to identify genes required for RNA interference. *Proc. Natl Acad. Sci. USA* **99**, 4191–4196 (2002).
27. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75**, 855–862 (1993).

28. Reinhart, B. J. *et al.* The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901–906 (2000).
29. Kennerdell, J. R. & Carthew, R. W. Use of dsRNA-mediated genetic interference to demonstrate that *frizzled 2* act in the wingless pathway. *Cell* **95**, 1017–1026 (1998).
30. Tuschl, T., Zamore, P. D., Lehmann, R., Bartel, D. P. & Sharp, P. A. Targeted mRNA degradation by double-stranded RNA in vitro. *Genes Dev.* **13**, 3191–3197 (1999).
31. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293–296 (2000).
32. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**, 950–952 (1999).
33. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).
34. Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W. & Tuschl, T. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* **20**, 6877–6888 (2001).
35. Filipov, V., Solov'yev, V., Filippova, M. & Gill, S. S. A novel type of RNase III family proteins in eukaryotes. *Gene* **245**, 213–221 (2000).
36. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366 (2001).
37. Ketting, R. F. *et al.* Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev.* **15**, 2654–2659 (2001).
38. Knight, S. W. & Bass, B. L. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. *Science* **293**, 2269–2271 (2001).
39. Grishok, A. *et al.* Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**, 23–34 (2001).
40. Blaszczyk, J. *et al.* Crystallographic and modeling studies of RNase III suggest a mechanism for double-stranded RNA cleavage. *Structure (Camb.)* **9**, 1225–1236 (2001).
41. Nykanen, A., Haley, B. & Zamore, P. D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107**, 309–321 (2001).
42. Hammond, S. M., Boettcher, S., Caudy, A. A., Kobayashi, R. & Hannon, G. J. Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* **293**, 1146–1150 (2001).
43. Bohmert, K. *et al.* *AGO1* defines a novel locus of *Arabidopsis* controlling leaf development. *EMBO J.* **17**, 170–180 (1998).
44. Hammond, S. M., Caudy, A. A. & Hannon, G. J. Post-transcriptional gene silencing by double-stranded RNA. *Nature Rev. Genet.* **2**, 110–119 (2001).
45. Grishok, A., Tabara, H. & Mello, C. C. Genetic requirements for inheritance of RNAi in *C. elegans*. *Science* **287**, 2494–2497 (2000).
46. Ketting, R. F., Haverkamp, T. H., van Luenen, H. G. & Plasterk, R. H. *mut-7* of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. *Cell* **99**, 133–141 (1999).
47. Parrish, S. & Fire, A. Distinct roles for RDE-1 and RDE-4 during RNA interference in *Caenorhabditis elegans*. *RNA* **7**, 1397–1402 (2001).
48. Catalano, C., Azzalin, G., Macino, G. & Cogoni, C. Involvement of small RNAs and role of the *qde* genes in the gene silencing pathway in *Neurospora*. *Genes Dev.* **16**, 790–795 (2002).
49. Palauqui, J. C., Elmayan, T., Pollien, J. M. & Vaucheret, H. Systemic acquired silencing: transgene-specific post-transcriptional silencing is transmitted by grafting from silenced stocks to non-silenced scions. *EMBO J.* **16**, 4738–4745 (1997).
50. Sijen, T. *et al.* On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107**, 465–476 (2001).
51. Schiebel, W. *et al.* Isolation of an RNA-directed RNA polymerase-specific cDNA clone from tomato. *Plant Cell* **10**, 2087–2101 (1998).
52. Mourrain, P. *et al.* *Arabidopsis* SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell* **101**, 533–542 (2000).
53. Cogoni, C. & Macino, G. Gene silencing in *Neurospora crassa* requires a protein homologous to RNA-dependent RNA polymerase. *Nature* **399**, 166–169 (1999).
54. Smardon, A. *et al.* EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in *C. elegans*. *Curr. Biol.* **10**, 169–178 (2000).
55. Lipardi, C., Wei, Q. & Paterson, B. M. RNAi as random degradative PCR: siRNA primers convert mRNA into dsRNAs that are degraded to generate new siRNAs. *Cell* **107**, 297–307 (2001).
56. Fabian, E., Jones, L. & Baulcombe, D. C. Spreading of RNA targeting and DNA methylation in RNA silencing requires transcription of the target gene and a putative RNA dependent RNA polymerase. *Plant Cell* **14**, 857–867 (2002).
57. Voinnet, O., Vain, P., Angell, S. & Baulcombe, D. C. Systemic spread of sequence-specific transgene RNA degradation in plants is initiated by localized introduction of ectopic promoterless DNA. *Cell* **95**, 177–187 (1998).
58. Mallory, A. C. *et al.* HC-Pro suppression of transgene silencing eliminates the small RNAs but not transgene methylation or the mobile signal. *Plant Cell* **13**, 571–583 (2001).
59. Winston, W. M., Molodowitch, C. & Hunter, C. P. Systemic RNAi in *C. elegans* requires the putative transmembrane protein SID-1. *Science* **295**, 2456–2459 (2002).
60. Cogoni, C. & Macino, G. Posttranscriptional gene silencing in *Neurospora* by a RecQ DNA helicase. *Science* **286**, 2342–2344 (1999).
61. Wu-Scharf, D., Jeong, B., Zhang, C. & Cerutti, H. Transgene and transposon silencing in *Chlamydomonas reinhardtii* by a DEAH-box RNA helicase. *Science* **290**, 1159–1162 (2000).
62. Tijsterman, M., Ketting, R. F., Okihara, K. L., Sijen, T. & Plasterk, R. H. RNA helicase MUT-14-dependent gene silencing triggered in *C. elegans* by short antisense RNAs. *Science* **295**, 694–697 (2002).
63. Domeier, M. E. *et al.* A link between RNA interference and nonsense-mediated decay in *Caenorhabditis elegans*. *Science* **289**, 1928–1931 (2000).
64. Martienssen, R. A. & Colot, V. DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science* **293**, 1070–1074 (2001).
65. Furner, I. J., Sheikh, M. A. & Collett, C. E. Gene silencing and homology-dependent gene silencing in Arabidopsis: genetic modifiers and DNA methylation. *Genetics* **149**, 651–662 (1998).
66. Fagard, M., Boutet, S., Morel, J. B., Bellini, C. & Vaucheret, H. AGO1, QDE-2, and VADE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proc. Natl Acad. Sci. USA* **97**, 11650–11654 (2000).
67. Voinnet, O., Lederer, C. & Baulcombe, D. C. A viral movement protein prevents spread of the gene silencing signal in *Nicotiana benthamiana*. *Cell* **103**, 157–167 (2000).
68. Baulcombe, D. Viruses and gene silencing in plants. *Arch. Virol. Suppl.* **15**, 189–201 (1999).
69. Jacobsen, S. E., Running, M. P. & Meyerowitz, E. M. Disruption of an RNA helicase/RNase III gene in Arabidopsis causes unregulated cell division in floral meristems. *Development* **126**, 5231–5243 (1999).
70. Kataoka, Y., Takeichi, M. & Uemura, T. Developmental roles and molecular characterization of a *Drosophila* homologue of Arabidopsis Argonaute1, the founder of a novel gene superfamily. *Genes Cells* **6**, 313–325 (2001).
71. Cox, D. N. *et al.* A novel class of evolutionarily conserved genes defined by *piwi* are essential for stem cell self-renewal. *Genes Dev.* **12**, 3715–3727 (1998).
72. Morel, J. B. *et al.* Fertile hypomorphic ARGONAUTE (*ago1*) mutants impaired in post-transcriptional gene silencing and virus resistance. *Plant Cell* **14**, 629–639 (2002).
73. Moussian, B., Schoof, H., Haecker, A., Jurgens, G. & Laux, T. Role of the *ZWILLE* gene in the regulation of central shoot meristem cell fate during *Arabidopsis* embryogenesis. *EMBO J.* **17**, 1799–1809 (1998).
74. Hutvagner, G. *et al.* A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. *Science* **293**, 834–838 (2001).
75. Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993).
76. Olsen, P. H. & Ambros, V. The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* **216**, 671–680 (1999).
77. Slack, F. J. *et al.* The *lin-41* RBCC gene acts in the *C. elegans* heterochronic pathway between the *let-7* regulatory RNA and the LIN-29 transcription factor. *Mol. Cell* **5**, 659–669 (2000).
78. Ha, I., Wightman, B. & Ruvkun, G. A bulged *lin-4*/*lin-14* RNA duplex is sufficient for *Caenorhabditis elegans* *lin-14* temporal gradient formation. *Genes Dev.* **10**, 3041–3050 (1996).
79. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
80. Paddison, P. J., Caudy, A. A., Bernstein, E., Hannon, G. J. & Conklin, D. S. Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev.* **16**, 948–958 (2002).
81. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853–858 (2001).
82. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).
83. Lee, R. C. & Ambros, V. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294**, 862–864 (2001).
84. Mourelatos, Z. *et al.* miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.* **16**, 720–728 (2002).
85. Lai, E. C. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genet.* **30**, 363–364 (2002).
86. Wianny, F. & Zernicka-Goetz, M. Specific interference with gene function by double-stranded RNA in early mouse development. *Nature Cell Biol.* **2**, 70–75 (2000).
87. Svoboda, P., Stein, P., Hayashi, H. & Schultz, R. M. Selective reduction of dormant maternal mRNAs in mouse oocytes by RNA interference. *Development* **127**, 4147–4156 (2000).
88. Gil, J. & Esteban, M. Induction of apoptosis by the dsRNA-dependent protein kinase (PKR): mechanism of action. *Apoptosis* **5**, 107–114 (2000).
89. Clarke, P. A. & Mathews, M. B. Interactions between the double-stranded RNA binding motif and RNA: definition of the binding site for the interferon-induced protein kinase DAI (PKR) on adenovirus VA RNA. *RNA* **1**, 7–20 (1995).
90. Smith, N. A. *et al.* Total silencing by intron-spliced hairpin RNAs. *Nature* **407**, 319–320 (2000).
91. Tavernarakis, N., Wang, S. L., Dorovkov, M., Ryazanov, A. & Driscoll, M. Heritable and inducible genetic interference by double-stranded RNA encoded by transgenes. *Nature Genet.* **24**, 180–183 (2000).
92. Kennerdell, J. R. & Carthew, R. W. Heritable gene silencing in *Drosophila* using double-stranded RNA. *Nature Biotechnol.* **18**, 896–898 (2000).
93. LaCount, D. J., Bruse, S., Hill, K. L. & Donelson, J. E. Double-stranded RNA interference in *Trypanosoma brucei* using head-to-head promoters. *Mol. Biochem. Parasitol.* **111**, 67–76 (2000).
94. Shi, H. *et al.* Genetic interference in *Trypanosoma brucei* by heritable and inducible double-stranded RNA. *RNA* **6**, 1069–1076 (2000).
95. Wang, Z., Morris, J. C., Drew, M. E. & Englund, P. T. Inhibition of *Trypanosoma brucei* gene expression by RNA interference using an integratable vector with opposing T7 promoters. *J. Biol. Chem.* **275**, 40174–40179 (2000).
96. Paddison, P. J., Caudy, A. A. & Hannon, G. J. Stable suppression of gene expression by RNAi in mammalian cells. *Proc. Natl Acad. Sci. USA* **99**, 1443–1448 (2002).
97. Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **21**, 21 (2002).
98. Sui, G. *et al.* A DNA vector-based RNAi technology to suppress gene expression in mammalian cells. *Proc. Natl Acad. Sci. USA* **99**, 5515–5520 (2002).
99. Meissner, W., Rothfels, H., Schafer, B. & Seifart, K. Development of an inducible pol III transcription system essentially requiring a mutated form of the TATA-binding protein. *Nucleic Acids Res.* **29**, 1672–1682 (2001).
100. Ohkawa, J. & Taira, K. Control of the functional activity of an antisense RNA by a tetracycline-responsive derivative of the human U6 snRNA promoter. *Hum. Gene Ther.* **11**, 577–585 (2000).

Acknowledgements

I thank members of the Hannon laboratory for critical reading of the manuscript; J. Duffy for help in preparation of the figures; D. Baulcombe, M. Tijsterman, R. Carthew and S. Prasanth for providing the images for Fig. 1; fellow investigators who granted permission to discuss unpublished observations; and C. Mello and C. Sherr for providing motive and opportunity, respectively, for our early work on RNAi. G.J.H. is a Rita Allen Foundation scholar and is supported by an Innovator Award from the U.S. Army Breast Cancer Research Program. This work was supported in part by a grant from the NIH.

Emerging clinical applications of RNA

Bruce A. Sullenger & Eli Gilboa

Department of Surgery, Duke University Medical Center, Durham, North Carolina 27710, USA (e-mail: b.sullenger@cgct.duke.edu)

RNA is a versatile biological macromolecule that is crucial in mobilizing and interpreting our genetic information. It is not surprising then that researchers have sought to exploit the inherent properties of RNAs so as to interfere with or repair dysfunctional nucleic acids or proteins and to stimulate the production of therapeutic gene products in a variety of pathological situations. The first generation of the resulting RNA therapeutics are now being evaluated in clinical trials, raising significant interest in this emerging area of medical research.

The concept of using RNA molecules as therapeutic agents is relatively new, but has received increasing attention during the past decade. Much of this interest stems from a variety of basic scientific discoveries that underscore the seminal role of RNA molecules in the utilization of genetic instructions in all living systems and the versatility of these molecules in nature. RNA molecules can adopt a wide variety of conformations and perform a range of cellular functions. Certain RNAs fold to form catalytic centres, whereas others have structures that allow them to make specific RNA–RNA, RNA–DNA or RNA–protein interactions.

Such realizations have led translational researchers to attempt to exploit various facets of RNA biology and chemistry to combat human disease. Significant progress has been made towards this goal and the first RNA-based therapeutics are now being evaluated in clinical trials for the treatment of disorders ranging from cancer to infectious diseases. The therapeutic RNAs that have so far received the most attention can be grouped into four categories: gene inhibitors, gene amenders, protein inhibitors and immunostimulatory RNAs. Here we review the development of these various RNA-based approaches to therapy and provide an update on the progress towards moving this emerging class of molecular therapeutics into and through clinical evaluation. Finally, we will discuss the developmental difficulties that various RNA therapeutics still face and consider potential solutions to those problems.

RNA-mediated inhibition of gene expression

Regulation of gene expression by an RNA that is complementary to a target messenger RNA (mRNA) was first recognized as a naturally occurring process in prokaryotes¹. These complementary RNAs, termed antisense RNAs, specifically recognize their target transcripts by forming base pairs with them in a sequence-dependent manner. The formation of this RNA duplex is believed to lead to the degradation of the target RNA or the inhibition of its translation. The ability to inhibit specific genes after gene transfer of antisense expression cassettes was first demonstrated almost two decades ago in bacteria by Pestka *et al.*² and Coleman *et al.*³ and in eukaryotic cells by Izant and Wientraub⁴. Following these initial studies, numerous reports appeared that described the potential utility of antisense RNA for the inhibition of a wide array of genes in mammalian cells⁵. However, these studies also indicated that the efficacy of antisense-mediated gene inhibition was usually dependent on the presence of a considerable excess of antisense RNA to

target RNA in the cell⁵. Therefore methods were sought to express large quantities of antisense RNAs in transduced cells⁶ or to generate antisense RNAs that could destroy multiple target RNAs and thus reduce the need for an excess of the inhibitory RNA.

The discovery that certain RNAs can perform catalysis^{7,8} has led to the development of a class of therapeutic RNAs called *trans*-cleaving ribozymes. Such ribozymes bind substrate RNAs through base-pairing interactions, cleave the bound target RNA, release the cleavage products and are recycled so that they can repeat this process multiple times *in vitro* (Fig. 1a). The observation that such ribozymes can be repeatedly targeted to cleave virtually any pathogenic transcript *in vitro*^{9,10} led to much speculation about their potential therapeutic value *in vivo*^{11,12}. Much progress has been made towards assessing the potential utility of *trans*-cleaving ribozymes, with the hammerhead and hairpin ribozyme¹³ being the main focus of this translational effort (see review in this issue by Doudna and Cech, pages 222–228).

Several phase I and II clinical trials have been initiated using *trans*-cleaving ribozymes in a small number of patients with infectious diseases or cancer. In these studies the ribozymes have been delivered to the patients either by gene therapy methods or by direct injection of a synthetic ribozyme. The gene therapy-based trials have focused upon developing ribozyme-based treatments for individuals infected with the human immunodeficiency virus (HIV). Three separate groups have used retroviral vectors to introduce expression cassettes for anti-HIV ribozymes into CD4⁺ lymphocytes or CD34⁺ haematopoietic precursors *ex vivo* that have been taken from the infected patient or from an identical twin^{14–16} (Fig. 1b). The transduced cells are then infused into the patient and the engraftment and survival of the ribozyme-containing cells are monitored.

Initial results from these studies suggest that transfer of ribozyme-encoding genes to HIV-infected individuals is well tolerated and transduced cells can persist in the patient¹⁵. Moreover, preliminary reports suggest that anti-HIV ribozyme-containing cells may possess a transient survival advantage in the patient compared with cells transduced with a control vector¹⁵. Unfortunately, such studies also indicate that gene transfer into long-term progenitors has not been accomplished because transduced cells decrease to below detection by one year after infusion (J. J. Rossi, personal communication). Larger clinical trials must now be performed to evaluate the efficacy of anti-HIV ribozymes. Critical factors that will influence the success of such trials will be the development of gene-transfer systems

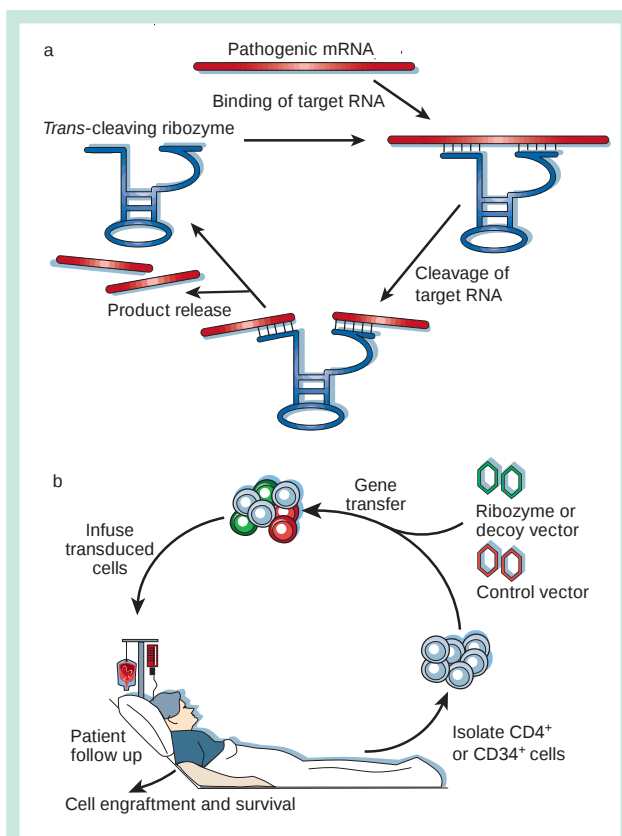


Figure 1 Applications of *trans*-cleaving ribozymes for gene inhibition. **a**, *Trans*-cleaving ribozymes can bind pathogenic target RNAs through base-pairing interactions, cleave the target, release the reaction products and repeat this process with multiple turnover. **b**, Clinical trials of HIV gene therapy with *trans*-cleaving ribozymes and decoy RNAs. Haematopoietic cells that express CD4 or CD34 on their surface are isolated from the patient and transduced with retroviral vectors containing an HIV inhibitory gene (ribzyme or decoy; green virus) or a control gene (red virus). Cells transduced with the vectors are re-introduced into the patient and the relative engraftment and survival of cells containing the HIV inhibitory gene (green) and control vector (red) are determined.

that can efficiently transduce pluripotent haematopoietic stem cells and the generation of improved ribozyme expression cassettes that can increase the survival advantage of transduced cells. In this regard, encouraging preclinical studies suggest that co-localization of ribozymes with their viral target RNAs inside cells may enhance ribozyme activity *in vivo*^{17,18}, and use of combinations of inhibitory ribozymes and decoy RNAs may yield more potent inhibitors of HIV against which it will be difficult for the virus to develop resistance.

Three different nuclease-resistant synthetic ribozymes are being evaluated in clinical trials¹². Each of these trials uses a hammerhead ribozyme derivative that contains chemical modifications that greatly increase the ribozyme's stability in biological fluids¹⁹. Moreover, methods have been developed that enable large-scale synthesis of this new class of therapeutic agent under good manufacturing practice protocols²⁰. All three of these synthetic ribozymes target RNAs whose expression is associated with the induction or progression of cancer and all three have shown promising results in preclinical cell and animal experiments^{21,22}.

In 1998, the first of these ribozymes entered a phase I trial targeting *flt-1* mRNA, which encodes the high-affinity receptor for the angiogenic protein vascular endothelial growth factor (VEGF). Results from this and two subsequent phase I trials show that daily

intravenous or subcutaneous delivery of this compound is well tolerated and that plasma levels could be maintained for prolonged periods after subcutaneous delivery¹². Currently its therapeutic efficacy is being evaluated in phase II trials for breast and colorectal cancer. The other two synthetic ribozymes to enter clinical trials target mRNA of human epidermal growth factor receptor type 2, which is overexpressed in many breast cancers, and hepatitis C virus (HCV) RNA, which is associated with liver cirrhosis and hepatocellular carcinoma. Results from these initial efficacy studies will provide the first significant insight into the long-term utility of *trans*-cleaving ribozymes as therapeutic agents.

The critical factors that will most likely determine the success of these synthetic ribozyme efficacy trials are the ability to deliver ribozymes efficiently into the appropriate cells *in vivo*, and the level and duration of target-gene inhibition that is required to alter disease pathophysiology and slow disease progression. For ribozymes to benefit individuals with chronic disorders such as cancer and HCV (or HIV) infection, long-term, high-level inhibition of target transcripts will probably be required. This may be difficult to achieve in practice, especially when targeting highly expressed viral RNAs. For example, we observed that a hammerhead ribozyme was able to inhibit the replication of a murine retrovirus by up to 90% by co-localizing the ribozyme with its viral target inside cells, but we found that higher levels of inhibition were difficult to achieve even when a large excess of the ribozyme was present in the cells¹⁷. Thus, the utility of *trans*-cleaving ribozymes may be limited to conditions where a modest reduction in pathogenic gene expression will result in therapeutic efficacy or to combination therapies where the ribozyme can act in concert with other therapeutics (for example, antiviral or chemotherapeutic agents) that largely control the pathogen.

Finally, two additional RNA-based strategies for gene inhibition in mammalian cells have recently been described. First, in many eukaryotic cells, expression of an mRNA can be inhibited by a double-stranded RNA that corresponds to the sequence of the targeted transcript. This process, known as RNA interference (RNAi; see review in this issue by Hannon, pages 244–251), has been shown to function in mammalian cells after introduction of short (~21 nucleotides), synthetic duplex RNAs²³ or expression of similar-length transcripts by RNA polymerase III (Pol III) promoters^{24–27}. Using this Pol III expression approach, Lee and colleagues²⁶ have shown that RNAi can inhibit HIV gene expression by up to 4 logs in co-transfection experiments. These and other results have raised much interest about the potential utility of the RNAi approach for therapeutic applications.

Second, in contrast to *trans*-cleaving ribozymes and RNAi, which target pathogenic RNA for destruction, mobile group II introns can be targeted to insert into and inactivate pathogenic DNA. The obvious advantage of targeting DNA rather than RNA is that such disruption would be a one-time event that would effectively knock out 100% of the RNAs issuing from the disrupted gene. Lambowitz and colleagues have shown²⁸ that a mobile group II intron can be retargeted to insert specifically into HIV pro-viral DNA or the HIV co-receptor gene *CCR5*. Gene disruption was efficient in bacteria where the HIV *pol* gene could be disrupted in over 60% of the cells. Group II mobilization into extra-chromosomal copies of the HIV *pol* and *CCR* genes in mammalian cells has been shown to be possible²⁸, but mobilization into genomic DNA has yet to be reported. Translational researchers must now begin to determine whether these two gene-inhibition strategies can be utilized for therapeutic applications. Information gleaned from the ongoing clinical trials evaluating *trans*-cleaving ribozymes will undoubtedly facilitate the clinical development of this next generation of RNA-based gene inhibitors.

RNA-mediated repair of genetic instructions

Genetic instructions are usually revised as they are converted from DNA to RNA to protein in human cells. Most of this revision occurs at

the RNA level when splicing removes intron sequences from precursor transcripts and ligates together flanking exon sequences to generate mature RNAs. Interestingly, many of the molecules involved in revising RNA messages are RNAs themselves. Several recent studies exploit this facet of RNA biology and describe the development of an intriguing new class of therapeutic RNAs that can perform *trans*-splicing to repair clinically relevant mutant transcripts. The concept of RNA repair has received much attention as a novel approach to gene therapy. Compared with more traditional strategies, RNA repair should engender the regulated expression of the corrected gene product and simultaneously reduce the expression of the mutant gene product. This property makes RNA repair an attractive strategy to attempt in the treatment of genetic disorders associated with mutant genes that are highly regulated or that encode deleterious mutant proteins.

The initial studies focused on RNA repair used a *trans*-splicing version of a group I ribozyme to repair mutant *lacZ* transcripts in bacteria²⁹ and mammalian cells³⁰. These studies showed that the ribozyme was able to repair the mutant RNA by recognizing the target transcript by base pairing with it, cleaving off mutant sequences and ligating a wild-type sequence onto the cleavage product (Fig. 2a). More recently, *trans*-splicing ribozymes have been generated that can amend mutant transcripts associated with myotonic dystrophy³¹ and many cancers (mutant *p53* tumour-suppressor transcripts)³² in mammalian cell lines, and with sickle cell anaemia in erythrocyte precursors isolated from patients with sickle cell disease³³.

A second, related approach to RNA repair uses spliceosomes to revise mutant transcripts by *trans*-splicing³⁴. In this case the spliceosome performs the *trans*-splicing reaction upon a pre-*trans*-splicing RNA molecule (termed a PTM) and a target RNA (Fig. 2b). An expression cassette for the PTM is delivered to the cell, whereas the spliceosome and target RNA are supplied by the cell. Puttaraju *et al.* first demonstrated that spliceosome-mediated RNA *trans*-splicing (SMaRT) could be used to reprogram human chorionic gonadotropin β -polypeptide mRNAs and to repair mutant *lacZ* transcripts in cell culture³⁵. Subsequently, encouraging results have shown that the SMaRT approach can repair a clinically relevant fraction of a mutant cystic fibrosis transmembrane conductance regulator (CFTR) transcript (CFTR Δ F508) in human cystic fibrosis airway-epithelial cells grown in culture or in animal xenografts³⁶. Such repair resulted in partial restoration of Cl⁻ transport in the CFTR-deficient cells to 12–15% of the level observed in wild-type CFTR-containing cells.

A critical question requiring further study for both ribozyme- and spliceosome-mediated repair of mutant RNAs is the specificity of RNA revision. The specificity of spliceosomal *trans*-splicing seems to be low, at least in certain instances, with the spliceosome *trans*-splicing the PTM sequence onto many unintended target RNAs in mammalian cells³⁷. Initial studies suggested that the specificity of ribozyme-mediated repair may also be low³⁰; but more recent investigations have described modifications to the ribozyme to address the issue of target specificity directly^{32,38–40}. Definitive studies are now warranted to address this question directly and to facilitate the development of more specific *trans*-splicing agents if needed. Even though some issues concerning RNA repair remain unresolved, the preclinical studies performed so far on therapeutic applications of *trans*-splicing are encouraging. They demonstrate that *trans*-splicing can amend mutant transcripts associated with a variety of human diseases and repair target RNAs with efficiencies that would be expected to be clinically beneficial, at least for treating many recessive disorders.

RNA as a protein antagonist

Many small RNAs can fold into three-dimensional structures that allow them to bind target proteins with high affinity and specificity. Several RNA viruses such as HIV use this property of RNA to recruit viral and host proteins to perform essential functions in viral

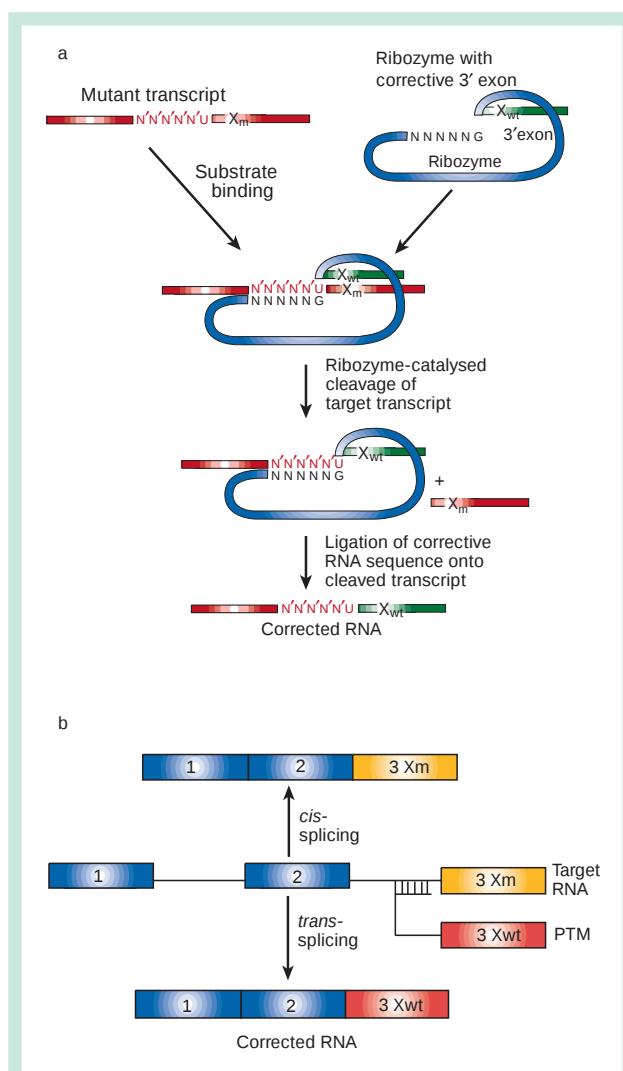
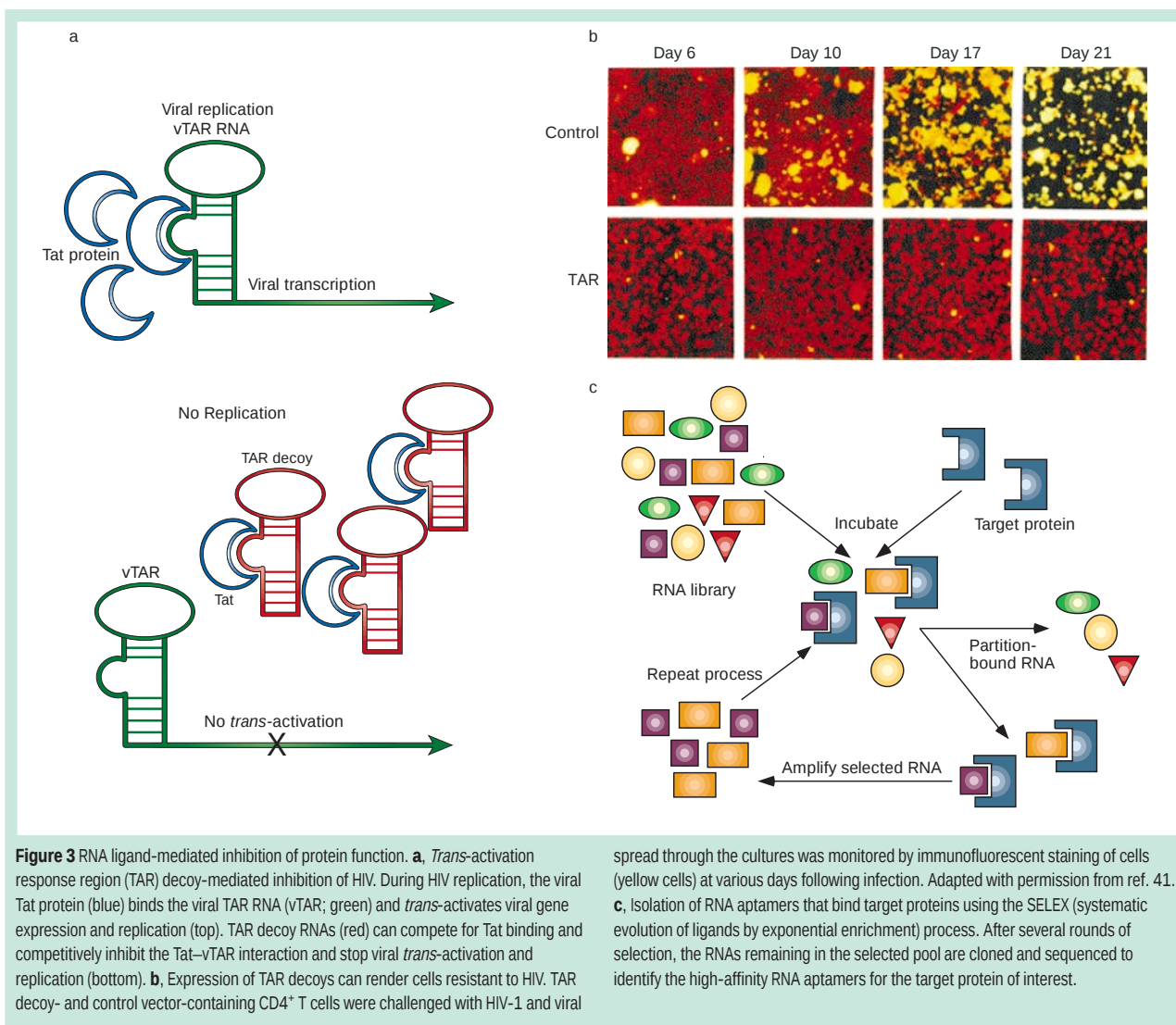


Figure 2 *Trans*-splicing-mediated repair of mutant transcripts. **a**, Ribozyme-mediated repair. *Trans*-splicing ribozymes recognize mutant RNAs upstream of a mutation site (Xm). The mutant RNA is cleaved and an exon with a wild-type sequence (Xwt) is ligated onto the cleavage product to generate a corrected transcript. **b**, Spliceosome-mediated RNA *trans*-splicing repair of a mutant transcript. A target RNA containing a mutation in its third exon (Xm) normally undergoes *cis*-splicing to generate an mRNA that encodes a defective gene product. A pre-*trans*-splicing RNA molecule (PTM, red) with a wild-type exon 3 (Xwt) can impede the *cis*-splicing process and engender *trans*-splicing to yield an mRNA with a correct sequence.

replication. For example, HIV uses small-structured RNA elements termed the *trans*-activation response region (TAR) and Rev response element (RRE) to recruit the viral regulatory proteins Tat and Rev to control viral gene expression. The use of small-structured RNAs to directly bind and inhibit the activity of a pathogenic protein was first explored using the HIV TAR sequence (Fig. 3a). Expression of TAR 'decoy' RNAs in CD4⁺ T cells was shown to competitively inhibit Tat binding to the viral TAR RNA and render cells highly resistant to HIV replication (Fig. 3b)⁴¹. Subsequent studies demonstrated that RRE also could act as a decoy RNA to block Rev activity and inhibit HIV replication⁴².

These observations suggested that short RNA decoys might be useful therapeutic agents for inhibiting HIV replication *in vivo* and have led to a phase I gene therapy-based clinical trial designed to test safety and feasibility. In this trial, retroviral vectors were used to



introduce expression cassettes for RRE decoys into haematopoietic progenitors that had been isolated from, and re-infused into, HIV-infected paediatric patients⁴³ (Fig. 1b). This initial study showed that gene transfer and expression of RRE decoys was well tolerated, but as with other HIV gene-therapy trials, the results emphasize the need for improved gene-transfer techniques to transduce pluripotent haematopoietic progenitors⁴³.

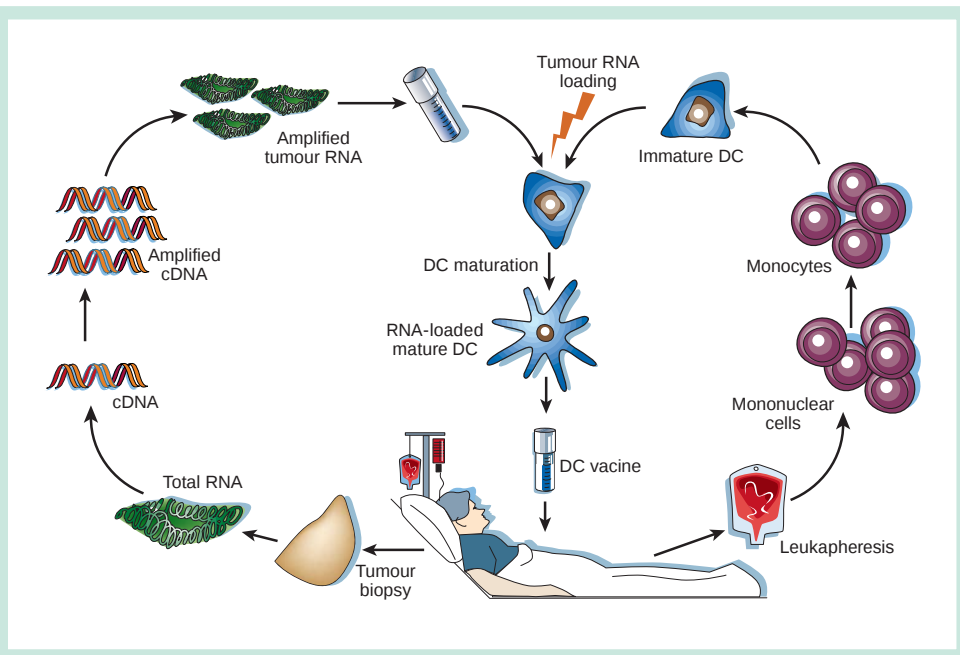
The observation that TAR and RRE decoy RNAs could be used to competitively inhibit viral protein function and replication suggested that other small-structured RNA molecules might be able to bind pathogenic target proteins and inhibit their activity. Concurrent with this observation, work from two groups suggested that iterative *in vitro* selection methods could be used to isolate high-affinity RNA ligands from large pools of randomized RNA sequences (vast RNA shape libraries) that could bind to proteins and small molecules^{44,45}. The resulting RNA ligands were termed aptamers by Ellington and Szostak⁴⁵ and the selection process was named SELEX (systematic evolution of ligands by exponential enrichment) by Tuerk and Gold⁴⁴ (Fig. 3c). SELEX has now been used to identify RNA aptamers that can bind to and inhibit the activity of a wide variety of proteins (for reviews, see refs 46, 47). The affinities of aptamers for their targets are similar to the affinities achieved by monoclonal antibodies for their antigens (K_d values typically in the low nanomolar to high picomolar range)⁴⁶. In contrast to antibodies, however, aptamers can be

spread through the cultures was monitored by immunofluorescent staining of cells (yellow cells) at various days following infection. Adapted with permission from ref. 41. **c**, Isolation of RNA aptamers that bind target proteins using the SELEX (systematic evolution of ligands by exponential enrichment) process. After several rounds of selection, the RNAs remaining in the selected pool are cloned and sequenced to identify the high-affinity RNA aptamers for the target protein of interest.

chemically synthesized to produce large quantities of these compounds for *in vivo* experimentation and clinical trials. Moreover, as with *trans*-cleaving ribozymes, synthetic aptamers can be modified to have greatly enhanced plasma stability and circulating half-lives and seem to exhibit low toxicity and immunogenicity *in vivo*^{47–49}.

So far only a few aptamers have been evaluated in animal models of disease (reviewed in refs 47, 50) and two of these have received the most attention. A DNA aptamer to thrombin functions as a potent anticoagulant that is able to maintain the patency of an extracorporeal circuit in sheep and replace heparin in a canine cardiopulmonary bypass model^{51–53}. A 2'-fluoro-modified RNA aptamer to VEGF-165 is able to inhibit neovascularization in a rat corneal-pocket angiogenesis assay and block glomerular endothelial cell proliferation and apoptosis in rats with mesangioproliferative nephritis^{49,54}. This VEGF aptamer is also the first therapeutic aptamer to be administered to humans and is currently being evaluated in a phase II/III clinical trial in patients with age-related macular degeneration. In this instance, the synthetic aptamer is being injected directly into the vitreous humour of the eye to assess treatment safety and efficacy. Factors that are likely to influence the trial success include the retention time of the aptamer in the vitreous humour and the relative importance of VEGF-165 on the progression of age-related macular degeneration in these patients. The observation that the VEGF aptamer remains active and can be recovered from the

Figure 4 Treatment of cancer patients with tumour RNA-transfected dendritic cells (DCs). In an outpatient setting, a tumour sample is removed from the patient and used to generate, and if necessary amplify, tumour RNA. Blood cells are obtained from the patient by leukapheresis and immature DCs are generated by culturing monocytes for 5–6 days in the presence of cytokines. Immature DCs are transfected with RNA and cultured another 24 hours in the presence of additional cytokines to mature. The antigen-loaded DCs can be cryopreserved for subsequent infusions into the patient.



vitreous humour of rhesus monkeys 28 days after injection⁵⁵ is encouraging with regard to prospects for long-term ocular retention of the compound.

Because most existing drugs are protein antagonists, the ultimate success of aptamers as therapeutic agents will probably depend on how well they compete with other classes of therapeutic compounds. In particular, the use of monoclonal antibodies has garnered increasing interest from both physicians and the pharmaceutical industry. Monoclonal antibodies are being generated and tested against many of the same proteins and for similar indications as most of the aptamers that are in pre-clinical development. Moreover, several antibodies have already progressed to market and are being used for the treatment of a variety of disorders. Therefore, in this competitive landscape, properties of aptamers that distinguish them from monoclonal antibodies (and other protein inhibitors) will have to be exploited for aptamers to penetrate the therapeutic market and ultimately fulfil their potential and become broadly useful pharmaceutical agents.

Immunotherapy using mRNA-transfected dendritic cells

Specific active immunotherapy of cancer — stimulating the patient's immune system to recognize and eliminate tumour cells — is emerging as a promising modality for treating cancer recurrence and low-volume metastatic disease. There is considerable evidence that the cytotoxic T lymphocyte (CTL) arm of the immune response is crucial in controlling tumour growth. CTLs recognize short (8–10 amino acids) antigenic peptides in association with major histocompatibility complex (MHC) molecules displayed on the cell surface. The peptides are generated in the cytoplasm via the proteolytic action of the multi-unit proteolytic complex, the proteasome. They are shuttled from the cytoplasm to the endoplasmic reticulum, where they associate with nascent MHC molecules, and are then transported to the cell surface for recognition by CTLs⁵⁶.

Naive CTLs generated in the thymus undergo an activation process to acquire the ability to kill their targets or secrete so-called 'effector' cytokines such as interferon- γ . Bone marrow-derived dendritic cells (DCs) displaying the appropriate MHC–peptide complex on the cell surface are the primary cell type capable of activating naive CTLs⁵⁷. Once activated, the CTL can recognize and kill any somatic cell presenting the MHC–peptide complex. Thus, to stimulate a CTL response against the tumour, the tumour antigens have to reach the

cytoplasm of a DC. In the cancer patient the capture of tumour antigens by DCs and stimulation of tumour-specific CTLs is presumed to be inefficient and a limiting factor in stimulating protective immunity.

One approach to stimulate effective CTL responses in cancer patients would be to reconstruct this process *in vitro*, that is, to isolate DCs from the patient, load them with tumour antigens in such a manner that the antigens reach the cytoplasm, and inject the antigen-loaded cells back into the patient. This approach has been accomplished by incubating DCs with peptides and proteins, or by transfecting the cells with DNA constructs. Transfecting DCs with mRNA-encoding (tumour) antigens is yet another way to load DCs with antigens⁵⁸. mRNA can be isolated directly from tumour cells or synthesized *in vitro* from complementary DNA templates. DCs transfected with mRNA-encoding specific antigens or total tumour-derived RNA elicited potent CTL responses and tumour immunity in mice^{59–64}, and DCs generated from healthy volunteers or from cancer patients transfected with tumour RNA stimulated CTL responses in culture^{62,64–74}.

Initial studies have used cationic lipids to facilitate the uptake of RNA by DCs^{58,63}. Remarkably, incubation of DCs with RNA alone was also sufficient to sensitize them to stimulate CTLs^{70–72,74}. This is despite the low efficiency of RNA transfer (measured by gene expression), and no doubt reflects the sensitivity of the immune system to recognize minute amounts of antigens not detectable by conventional means. Recently an improved method for DC transfection with RNA was developed using electroporation, which rivals the best transfection methods for nucleic acids^{66,69}.

Just how efficient is the mRNA transfection protocol? Using functional end points such as CTL priming or induction of tumour immunity in mice, studies comparing several loading techniques — including loading DCs with peptides and proteins, transfection with cDNA plasmids or transduction with vaccinia vectors — found that mRNA loading was invariably superior^{65,66,68,69}. Additionally, it is often not fully appreciated that the generation of mRNA-encoding specific antigens is a simple process. Given the sequence, a cDNA template can be generated from the cell by reverse transcription followed by amplification using the polymerase chain reaction and transcribed into RNA in a matter of a few hours. This can be compared to the task of generating the corresponding protein or identifying the antigenic peptides.

Use of mRNA-encoded antigens for cancer vaccination offers another potentially useful benefit. Only a handful of tumour antigens have been discovered so far, mostly from melanoma patients, and there is growing evidence that many of these antigens are not well suited for vaccination^{75,76}. The alternative option is to vaccinate with tumour-derived antigenic mixtures, an approach which animal studies suggest is surprisingly effective. In reality, however, it is not possible to obtain sufficient tumour tissue from most cancer patients to generate the amount of antigens thought necessary for an effective vaccination protocol. Consequently, a significant proportion (perhaps most) of cancer patients would not benefit from current vaccination strategies, regardless of how effective they might be. Use of mRNA as source of antigen offers a solution that has been shown to work^{62,71}. Biologically active RNA can be amplified from microscopic amounts of tumour tissue (for instance, from frozen sections or from needle biopsies) to provide a virtually inexhaustible amount of antigen from practically every patient.

Perhaps no less important is that use of mRNA-encoded antigens can address the concern associated with the emergence of treatment-resistant tumour variants which lost the antigens targeted by vaccination, a common occurrence in cancer therapy. Using amplification protocols, sufficient RNA can be readily generated from a microscopic amount of a newly emerging antigen-loss tumour variant, before it is too late.

In summary, the preclinical experience suggests that cancer vaccination with tumour RNA-transfected DCs may constitute a highly effective and broadly applicable treatment for patients with recurring cancer. Whether such observations can be translated to the clinic remains to be seen, but hints from initial clinical trials are not discouraging.

The primary drawback of DC-based vaccination is that it is a customized form of cell therapy; DCs (and tumour RNA) are generated from each patient and require *in vitro* manipulation of the cells before re-injection into the patient (Fig. 4). This adds cost and complexity to the treatment and remains a challenge owing to the yet unproven nature of this new paradigm in human therapy. The expectation, which we believe is well founded, is that improved outcome to otherwise intractable diseases will offset the added cost and complexity associated with this form of therapy. A phase I clinical trial in patients with metastatic prostate cancer vaccinated with prostate-specific antigen (PSA) RNA-transfected DCs (PSA is a common tumour-associated antigen in patients with prostate cancer) was recently concluded⁷⁷. The DC therapy seems to be safe and, despite the advanced state of the disease, all patients have responded immunologically to the vaccine, that is, they have all exhibited PSA-specific T-cell responses. Surprisingly, six out of seven available patients exhibited a (very) modest clinically related response, a small but statistically significant impact on the blood PSA levels. This modest effect is unlikely to translate to clinical benefit to the patients, but suggests that the approach of vaccination with RNA-loaded DCs deserves further consideration. A second trial in renal cancer patients vaccinated with tumour RNA-transfected DCs was recently completed and is being evaluated.

Perspectives

In the past five years, a number of clinical trials have been initiated to begin to evaluate the safety and efficacy of a variety of innovative RNA-based therapeutic strategies. RNA therapeutics that can inhibit gene expression, block protein function or induce potent immune (CTL) responses have all entered clinical trials. Moreover, the RNA therapeutic developmental pipeline is burgeoning and the next generation of RNA-based therapies is quickly making its way through pre-clinical studies. The ultimate clinical utility of any of these treatment modalities is obviously still unclear. Nevertheless, the impressive diversity of therapeutic RNAs that are being developed and the breadth of clinical indications that one can foresee treating with this new class of therapeutic agents is remarkable. Just as nature

has evolved elegant strategies that utilize RNA molecules to perform a variety of essential biological functions for the initiation and maintenance of life, we are confident that clinical researchers will develop innovative strategies that harness the utility of RNA molecules for the protection and improvement of human life. □

- Green, P. J., Pines, O. & Inouye, M. The role of antisense RNA in gene regulation. *Annu. Rev. Biochem.* **55**, 569–597 (1986).
- Pestka, S., Daugherty, B. L., Jung, V., Hotta, K. & Pestka, R. K. Anti-mRNA: specific inhibition of translation of single mRNA molecules. *Proc. Natl Acad. Sci. USA* **81**, 7525–7528 (1984).
- Coleman, J., Green, P. J. & Inouye, M. The use of RNAs complementary to specific mRNAs to regulate the expression of individual bacterial genes. *Cell* **37**, 429–436 (1984).
- Izant, J. G. & Weintraub, H. Constitutive and conditional suppression of exogenous and endogenous genes by anti-sense RNA. *Science* **229**, 345–352 (1985).
- van der Krol, A. R., Mol, J. N. & Stuitje, A. R. Modulation of eukaryotic gene expression by complementary RNA or DNA sequences. *Biotechniques* **6**, 958–976 (1988).
- Sullenger, B. A., Lee, T. C., Smith, C. A., Ungers, G. E. & Gilboa, E. Expression of chimeric tRNA-driven antisense transcripts renders NIH 3T3 cells highly resistant to Moloney murine leukemia virus replication. *Mol. Cell. Biol.* **10**, 6512–6523 (1990).
- Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147–157 (1982).
- Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–857 (1983).
- Uhlenbeck, O. C. A small catalytic oligoribonucleotide. *Nature* **328**, 596–600 (1987).
- Haseloff, J. & Gerlach, W. L. Simple RNA enzymes with new and highly specific endoribonuclease activities. *Nature* **334**, 585–591 (1988).
- Cech, T. R. Ribozymes and their medical implications. *J. Am. Med. Assoc.* **260**, 3030–3034 (1988).
- Usman, N. & Blatt, L. M. Nuclease-resistant synthetic ribozymes: developing a new class of therapeutics. *J. Clin. Invest.* **106**, 1197–1202 (2000).
- Symons, R. H. Small catalytic RNAs. *Annu. Rev. Biochem.* **61**, 641–671 (1992).
- Bauer, G. *et al.* Inhibition of human immunodeficiency virus-1 (HIV-1) replication after transduction of granulocyte colony-stimulating factor-mobilized CD34+ cells from HIV-1-infected donors using retroviral vectors containing anti-HIV-1 genes. *Blood* **89**, 2259–2267 (1997).
- Wong-Staal, F., Poeschla, E. M. & Looney, D. J. A controlled, Phase 1 clinical trial to evaluate the safety and effects in HIV-1 infected humans of autologous lymphocytes transduced with a ribozyme that cleaves HIV-1 RNA. *Hum. Gene Ther.* **9**, 2407–2425 (1998).
- Amado, R. G. *et al.* A phase I trial of autologous CD34+ hematopoietic progenitor cells transduced with an anti-HIV ribozyme. *Hum. Gene Ther.* **10**, 2255–2270 (1999).
- Sullenger, B. A. & Cech, T. R. Tethering ribozymes to a retroviral packaging signal for destruction of viral RNA. *Science* **262**, 1566–1569 (1993).
- Lee, N. S., Bertrand, E. & Rossi, J. mRNA localization signals can enhance the intracellular effectiveness of hammerhead ribozymes. *RNA* **5**, 1200–1209 (1999).
- Beigelman, L. *et al.* Chemical modification of hammerhead ribozymes. Catalytic activity and nuclease resistance. *J. Biol. Chem.* **270**, 25702–25708 (1995).
- Wincott, F. *et al.* Synthesis, deprotection, analysis and purification of RNA and ribozymes. *Nucleic Acids Res.* **23**, 2677–2684 (1995).
- Pavco, P. A. *et al.* Antitumor and antimetastatic activity of ribozymes targeting the messenger RNA of vascular endothelial growth factor receptors. *Clin. Cancer Res.* **6**, 2094–2103 (2000).
- Macejak, D. G. *et al.* Inhibition of hepatitis C virus (HCV)-RNA-dependent translation and replication of a chimeric HCV poliovirus using synthetic stabilized ribozymes. *Hepatology* **31**, 769–776 (2000).
- Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
- Sui, G. *et al.* A DNA vector-based RNAi technology to suppress gene expression in mammalian cells. *Proc. Natl Acad. Sci. USA* **99**, 5515–5520 (2002).
- Miyagishi, M. & Taira, K. U6 promoter-driven siRNA with four uridine 3' overhangs efficiently suppress targeted gene expression in mammalian cells. *Nature Biotechnol.* **20**, 497–500 (2002).
- Lee, N. S. *et al.* Expression of small interfering RNAs targeted against HIV-1 rev transcripts in human cells. *Nature Biotechnol.* **20**, 500–505 (2002).
- Paul, C. P. *et al.* Effective expression of small interfering RNA in human cells. *Nature Biotechnol.* **20**, 505–508 (2002).
- Guo, H. *et al.* Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* **289**, 452–457 (2000).
- Sullenger, B. A. & Cech, T. R. Ribozyme-mediated repair of defective mRNA by targeted, trans-splicing. *Nature* **371**, 619–622 (1994).
- Jones, J. T., Lee, S. W. & Sullenger, B. A. Tagging ribozyme reaction sites to follow trans-splicing in mammalian cells. *Nature Med.* **2**, 643–648 (1996).
- Phylactou, L. A., Darrah, C. & Wood, M. J. Ribozyme-mediated trans-splicing of a trinucleotide repeat. *Nature Genet.* **18**, 378–381 (1998).
- Watanabe, T. & Sullenger, B. A. Induction of wild-type p53 activity in human cancer cells by ribozymes that repair mutant p53 transcripts. *Proc. Natl Acad. Sci. USA* **97**, 8490–8494 (2000).
- Lan, N., Howrey, R. P., Lee, S. W., Smith, C. A. & Sullenger, B. A. Ribozyme-mediated repair of sickle beta-globin mRNAs in erythrocyte precursors. *Science* **280**, 1593–1596 (1998).
- Puttaraju, M., Jamison, S. F., Mansfield, S. G., Garcia-Blanco, M. A. & Mitchell, L. G. Spliceosome-mediated RNA trans-splicing as a tool for gene therapy. *Nature Biotechnol.* **17**, 246–252 (1999).
- Puttaraju, M., DiPasquale, J., Baker, C. C., Mitchell, L. G. & Garcia-Blanco, M. A. Messenger RNA repair and restoration of protein function by spliceosome-mediated RNA trans-splicing. *Mol. Ther.* **4**, 105–114 (2001).
- Liu, X. *et al.* Partial correction of endogenous ΔF508 CFTR in human cystic fibrosis airway epithelia by spliceosome-mediated RNA trans-splicing. *Nature Biotechnol.* **20**, 47–52 (2002).
- Kikumori, T., Cote, G. J. & Gagel, R. F. Promiscuity of pre-mRNA spliceosome-mediated trans splicing: a problem for gene therapy? *Hum. Gene Ther.* **12**, 1429–1441 (2001).
- Kohler, U., Ayre, B. G., Goodman, H. M. & Haseloff, J. Trans-splicing ribozymes for targeted gene delivery. *J. Mol. Biol.* **285**, 1935–1950 (1999).

39. Ayre, B. G., Kohler, U., Goodman, H. M. & Haseloff, J. Design of highly specific cytotoxins by using trans-splicing ribozymes. *Proc. Natl Acad. Sci. USA* **96**, 3507–3512 (1999).
40. Zarrinkar, P. P. & Sullenger, B. A. Optimizing the substrate specificity of a group I intron ribozyme. *Biochemistry* **38**, 3426–3432 (1999).
41. Sullenger, B. A., Gallardo, H. F., Ungers, G. E. & Gilboa, E. Overexpression of TAR sequences renders cells resistant to human immunodeficiency virus replication. *Cell* **63**, 601–608 (1990).
42. Lee, T. C., Sullenger, B. A., Gallardo, H. F., Ungers, G. E. & Gilboa, E. Overexpression of RRE-derived sequences inhibits HIV-1 replication in CEM cells. *New Biol.* **4**, 66–74 (1992).
43. Kohn, D. B. *et al.* A clinical trial of retroviral-mediated transfer of a rev-responsive element decoy gene into CD34⁺ cells from the bone marrow of human immunodeficiency virus-1-infected children. *Blood* **94**, 368–371 (1999).
44. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
45. Ellington, A. D. & Szostak, J. W. *In vitro* selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
46. Gold, L., Polisky, B., Uhlenbeck, O. & Yarus, M. Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* **64**, 763–797 (1995).
47. White, R. R., Sullenger, B. A. & Rusconi, C. P. Developing aptamers into therapeutics. *J. Clin. Invest.* **106**, 929–934 (2000).
48. Jellinek, D. *et al.* Potent 2'-amino-2'-deoxypyrimidine RNA inhibitors of basic fibroblast growth factor. *Biochemistry* **34**, 11363–11372 (1995).
49. Ruckman, J. *et al.* 2'-Fluoropyrimidine RNA-based aptamers to the 165-amino acid form of vascular endothelial growth factor (VEGF165). Inhibition of receptor binding and VEGF-induced vascular permeability through interactions requiring the exon 7-encoded domain. *J. Biol. Chem.* **273**, 20556–20567 (1998).
50. Hicke, B. J. & Stephens, A. W. Escort aptamers: a delivery service for diagnosis and therapy. *J. Clin. Invest.* **106**, 923–928 (2000).
51. Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H. & Toole, J. J. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature* **355**, 564–566 (1992).
52. Griffin, L. C., Tidmarsh, G. F., Bock, L. C., Toole, J. J. & Leung, L. L. In vivo anticoagulant properties of a novel nucleotide-based thrombin inhibitor and demonstration of regional anticoagulation in extracorporeal circuits. *Blood* **81**, 3271–3276 (1993).
53. DeAnda, A. Jr *et al.* Pilot study of the efficacy of a thrombin inhibitor for use during cardiopulmonary bypass. *Ann. Thoracic Surg.* **58**, 344–350 (1994).
54. Ostendorf, T. *et al.* Specific antagonism of PDGF prevents renal scarring in experimental glomerulonephritis. *J. Am. Soc. Nephrol.* **12**, 909–918 (2001).
55. Drolet, D. W. *et al.* Pharmacokinetics and safety of an anti-vascular endothelial growth factor aptamer (NX1838) following injection into the vitreous humor of rhesus monkeys. *Pharm. Res.* **17**, 1503–1510 (2000).
56. Yewdell, J. W., Norbury, C. C. & Bennink, J. R. Mechanisms of exogenous antigen presentation by MHC class I molecules in vitro and in vivo: implications for generating CD8⁺ T cell responses to infectious agents, tumors, transplants, and vaccines. *Adv. Immunol.* **73**, 1–77 (1999).
57. Banchereau, J. & Steinman, R. M. Dendritic cells and the control of immunity. *Nature* **392**, 245–252 (1998).
58. Boczkowski, D., Nair, S. K., Snyder, D. & Gilboa, E. Dendritic cells pulsed with RNA are potent antigen-presenting cells in vitro and in vivo. *J. Exp. Med.* **184**, 465–472 (1996).
59. Koido, S. *et al.* Induction of antitumor immunity by vaccination of dendritic cells transfected with MUC1 RNA. *J. Immunol.* **165**, 5713–5719 (2000).
60. Granstein, R. D., Ding, W. & Ozawa, H. Induction of anti-tumor immunity with epidermal cells pulsed with tumor-derived RNA or intradermal administration of RNA. *J. Invest. Dermatol.* **114**, 632–636 (2000).
61. Zhang, W. *et al.* Enhanced therapeutic efficacy of tumor RNA-pulsed dendritic cells after genetic modification with lymphotactin. *Hum. Gene Ther.* **10**, 1151–1161 (1999).
62. Boczkowski, D., Nair, S. K., Nam, J. H., Lyster, H. K., & Gilboa, E. Induction of tumor immunity and cytotoxic T lymphocyte responses using dendritic cells transfected with messenger RNA amplified from tumor cells. *Cancer Res.* **60**, 1028–1034 (2000).
63. Ashley, D. M. *et al.* Bone marrow-generated dendritic cells pulsed with tumor extracts or tumor RNA induce antitumor immunity against central nervous system tumors. *J. Exp. Med.* **186**, 1177–1182 (1997).
64. Nair, S. K. *et al.* Induction of cytotoxic T cell responses and tumor immunity against unrelated tumors using telomerase reverse transcriptase RNA transfected dendritic cells. *Nature Med.* **6**, 1011–1017 (2000).
65. Weissman, D. *et al.* HIV gag mRNA transfection of dendritic cells (DC) delivers encoded antigen to MHC class I and II molecules, causes DC maturation, and induces a potent human in vitro primary immune response. *J. Immunol.* **165**, 4710–4717 (2000).
66. Van Tendeloo, V. *et al.* Highly efficient gene delivery by mRNA electroporation in human hematopoietic cells: superiority to lipofection and passive pulsing of mRNA and to electroporation of plasmid cDNA for tumor antigen loading of dendritic cells. *Blood* **98**, 49–56 (2001).
67. Su, Z., Peluso, M. V., Raffegerst, S. H., Schendel, D. J. & Roskrow, M. A. The generation of LMP2a-specific cytotoxic T lymphocytes for the treatment of patients with Epstein-Barr virus-positive Hodgkin disease. *Eur. J. Immunol.* **31**, 947–958 (2001).
68. Strobel, I. *et al.* Human dendritic cells transfected with either RNA or DNA encoding influenza matrix protein M1 differ in their ability to stimulate cytotoxic T lymphocytes. *Gene Ther.* **7**, 2028–2035 (2000).
69. Saeboe-Larsen, S., Fossberg, E. & Gaudernack, G. mRNA-based electrotransfection of human dendritic cells and induction of cytotoxic T lymphocyte responses against the telomerase catalytic subunit (hTERT). *J. Immunol. Meth.* **259**, 191–203 (2002).
70. Heiser, A. *et al.* Human dendritic cells transfected with renal tumor RNA stimulate polyclonal T-cell responses against antigens expressed by primary and metastatic tumors. *Cancer Res.* **61**, 3388–3393 (2001).
71. Heiser, A. *et al.* Induction of polyclonal prostate cancer-specific CTL using dendritic cells transfected with amplified tumor RNA. *J. Immunol.* **166**, 2953–2960 (2001).
72. Heiser, A. *et al.* Human dendritic cells transfected with RNA encoding prostate-specific antigen stimulate prostate-specific CTL responses in vitro. *J. Immunol.* **164**, 5508–5514 (2000).
73. Thornburg, C., Boczkowski, D., Gilboa, E. & Nair, S. K. Induction of cytotoxic T lymphocytes with dendritic cells transfected with human papillomavirus E6 and E7 RNA: implications for cervical cancer immunotherapy. *J. Immunother.* **23**, 412–418 (2000).
74. Nair, S. K. *et al.* Induction of primary carcinoembryonic antigen (CEA)-specific cytotoxic T lymphocytes in vitro using human dendritic cells transfected with RNA. *Nature Biotechnol.* **16**, 364–369 (1998).
75. Srivastava, P. K. Do human cancers express shared protective antigens? Or the necessity of remembrance of things past. *Semin. Immunol.* **8**, 295–302 (1996).
76. Gilboa, E. The makings of a tumor rejection antigen. *Immunity* **11**, 263–270 (1999).
77. Heiser, A. *et al.* Autologous dendritic cells transfected with prostate-specific antigen RNA stimulate CTL responses against metastatic prostate tumors. *J. Clin. Invest.* **109**, 409–417 (2002).

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Senior European Sales Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)
Amelie Pequignot (4974)
Laura Pearson (4977)

Netherlands/ Italy/ Iberia:

Evelina Rubio Hakansson (4973)
Scandinavia: Sille Opstrup (4994)

Production Manager:

Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Ashly de Leon

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakata-machi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.johnson@naturejpn.com

Asia-Pacific Advertising Manager:

Kevyn Johnson

naturejobs

A chemical imbalance

At first glance, the results from the latest American Chemical Society (ACS) salary and employment survey seem curiously bifurcated. Compared with a year ago, unemployment has more than doubled among chemists seeking full-time work in a survey period ending this March. But salaries for those who have a job rose substantially more than the rate of inflation during the same period. This is a curious turn of events, as one would expect salaries to slip along with demand.

The ACS provides a partial explanation. It says that salaries increased, on average, by 4.8% this year (to \$76,600 for all chemists and \$82,500 for PhDs) because chemists were belatedly catching up with the economic boom that lasted from 1993 to 2000. But the organization doesn't delve into why unemployment rose from 1.5% last year to 3.3% this year. The answer could be even more simple: that the industry is feeling the effects of the current malaise in the general US economy.

A big part of that malaise might be a slowdown in pharmaceutical recruiting. At last year's ACS conference, career experts noted that the biggest area of job growth was organic chemistry, thanks to the increasing needs of biotechnology and drug companies for chemists who can assist with drug discovery and development. It stands to reason that, as those companies stumble, so too do the employment opportunities for chemists.

But the difficulty should be only temporary, as the number of newly qualified chemists entering the worldwide market, especially in the United States and Britain, is currently in steady decline (see *Nature* **416**, 777; 2002). This reduction in supply could be one reason why chemists' salaries have continued to rise — and why the downturn in job vacancies is likely to be short-lived.

Paul Smaglik

Naturejobs editor



Contents

POSTDOCS & STUDENTSHIPS

Making the transition from bench to bedside

p5

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS

Gaining a lot from translation

Steve O'Rahilly:
Dedicated clinical research facilities, which forge a link between bench and bedside, make it easier to pursue projects.



Converting basic science into clinical practice is never an easy task, and Europe is still lagging behind the United States in this 'bench-to-bedside' endeavour. The number of research centres in Europe that can offer dedicated space for 'translational research' is low, and funding for clinical scientists is difficult to obtain across the continent.

Unlike in the United States, where physicians can undergo scientific training during their education, research is rarely part of the curriculum in Europe. Germany's main research funding body, the DFG, admitted as such a few years ago in a statement, concluding that "research is still considered a spare time activity" at German academic medical centres.

In an effort to bridge the gap between basic and clinical research, the European Molecular Biology Laboratory in Heidelberg, Germany, teamed up with the University of Heidelberg Medical School and this January announced the opening of a Molecular Medicine Partnership Unit. Matthias Hentze, one of the unit's co-directors, hopes that the lab will grow to about 30 members for a pilot phase of four years. He sees the venture as an "incubator" for scientifically inclined physicians and would also like to initiate an MD/PhD programme.

FRESH FACILITIES

In the United Kingdom, the Wellcome Trust, the country's largest biomedical charity, is contributing £20 million (US\$30.5 million) to help to establish clinical research facilities (CRFs) in an effort to ease the burdens of hospitals that find it difficult to accommodate the needs of academic research. Five CRFs — the most recent of which opened in Manchester this month — have been set up as partnerships between local universities and local healthcare providers. The facilities will offer dedicated clinical space, equipment and personnel for several research labs.

For Helen Routledge, who is doing her PhD at the Queen Elizabeth Hospital in Birmingham, the CRF set up there in January is a major asset. She is investigating the effects of air pollution on the cardiovascular system. The improved access to equipment and the technical expertise provided by

the staff, which includes a cardiac technician, are a great help, she says.

Steve O'Rahilly, who heads the CRF opened at Addenbrooke's Hospital in Cambridge this January, says that the set-up makes it much easier to conduct his research into the genetic basis of severe obesity. As well as offering patients comfortable surroundings, he says, the in-patient facility means that study volunteers can stay overnight.

GROWING DEMAND

But despite this promise, the situation still requires some juggling skills. Sadaf Farooqi, a research fellow working for O'Rahilly, is pursuing bench research in addition to her clinical duties. Although she has managed to balance her responsibilities inside and outside the lab, she is quick to point out that her situation is not typical and that there are currently "not enough fellowships to meet the growing need". Farooqi believes it will be crucial to secure funding so that she can hire some help to keep her research afloat while she is on the wards honing her clinical skills.

Simon Hart, a specialist registrar in respiratory medicine at the University of Edinburgh, UK, has successfully taken the fellowship hurdle and secured a Clinician Scientist Fellowship from the Medical Research Council. The fellowship provides £6,500 a year for up to four years of support for the transition from postdoctoral fellow to independent investigator. Hart, whose research interest lies in the field of inflammatory-cell biology, is relieved that this type of funding does not force him to abandon his research project for the next four to five years while completing his medical training.

But some problems remain. Clinical scientists often face "financial penalties" for choosing research over practice, Hart says, and physicians tend to earn more than clinical researchers. Although the new CRFs are attracting more translational researchers, the salary gap may mean a continuing shortage of suitably trained candidates for academic clinical investigator positions. Unless new schemes emerge to make up the difference, it is likely that physician scientists will heed the call of clinical practice rather than translational research. ■

Jan Schmollinger is a graduate student in Boston.

Web links

Medical Research Council
♦ www.mrc.ac.uk
Wellcome Trust
♦ www.wellcome.ac.uk
European Molecular
Biology Laboratory
♦ www.embl-heidelberg.de